AD_____

Award Number: W81XWH-04-1-0570

TITLE: Endocrine Therapy of Breast Cancer

PRINCIPAL INVESTIGATOR: Robert Clarke, Ph.D., D.Sc.

CONTRACTING ORGANIZATION: Georgetown University
Washington, DC 20007

REPORT DATE: June 2007

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-06-2007	Annual	117 May 2006 – 16 May 2007

4. TITLE AND SUBTITLE

Endocrine Therapy of Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0570

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Robert Clarke, Ph.D., D.Sc.

Email: clarker@georgetown.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Georgetown University
Washington, DC 20007

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT A recent controversy in the treatment of estrogen receptor positive (ER+) breast cancers is whether an aromatase inhibitor, e.g., letrozole (LET) or TAM should be given as first line endocrine therapy. Unfortunately, response rates are lower, and response durations are shorter, on crossover than when these agents are given as first line therapies, e.g., ~40% of tumors show crossresistance to TAM or an aromatase inhibitor on crossover. Only 50% of ER+ tumors respond to endocrine therapy. Currently, we fail to predict endocrine responsiveness in about 66% of ER+/PgR- (progesterone receptor), 55% of ER-/PgR+, and 25% of ER+/PgR+ tumors. In this new Clinical Translational Research Award, we hypothesize that our analytical methods can extract expression profiles of breast tumors that define their responsiveness (sensitive vs. resistant) to endocrine therapy. These profiles, when combined with known predictive/prognostic factors, will support neural network and biostatistical classifiers or committee machines that predict each tumor's endocrine responsiveness. Our objectives are to array breast cancer cases, build classifiers of endocrine responsiveness (using microarray data), and validate these classifiers in independent data sets. In the long term, we will design custom arrays for use in clinical practice. Genes will be further studied using cellular and molecular methods, and their role as therapeutic targets explored.

15. SUBJECT TERMS Antiestrogen, aromatase inhibitor, anastrazole, bioinformatics, biomarkers, biostatistics, breast cancer, class prediction, clinical trial, computer science, engineering, immunohistochemistry, letrozole, microarrays, molecular profiling, neural networks, recurrence, resistance, tamoxifen.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	65	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

1. Ressom, et al. *IEEE Symp Compl Intel Bioinf Comput Biol*, 435-442, 2006.
2. Wang, et al. *Cancer Cell*, 10: 487-499, 2006.
3. Bouker, et al. *Cancer Genet Cytogenet*, 175: 61-4, 2007.
4. Wong, et al. *BMC Cancer* 6: 111-132, 2006.
5. Kuske, et al. *Endocr Related Cancer*, 13: 1121-1133, 2006.
6. Feng, et al. *6th IEEE Symp Bioinf Bioeng (BIBE '06)*, 165-170, 2006.
7. Gong, et al. *Proc 28th IEEE EMBS Intl Conf*, pp. 5872-5875, 2006.

INTRODUCTION

Endocrine therapy is often the least toxic and most effective treatment for hormone receptor positive invasive breast cancer. Such therapy includes antiestrogens (tamoxifen, fulvestrant) and aromatase inhibitors (anastrozole, letrozole, exemestane). Tamoxifen (TAM) increases disease free and overall survival in the adjuvant setting, reduces the incidence of estrogen receptor positive disease (ER+; unless otherwise noted ER=ERα) in high-risk women, and reduces the rate of bone loss secondary to osteoporosis in postmenopausal women [1,2]. Aromatase inhibitors are effective only in the absence of functioning ovaries - TAM can be used regardless of menopausal status. Recent studies suggest that anastrozole may be superior to TAM in the adjuvant treatment of postmenopausal women with ER+ breast cancer; other studies report higher overall response rates with letrozole (LET) *vs.* TAM as first line therapy in the metastatic setting. Thus, a recent controversy in the management of patients with ER+ disease is whether an aromatase inhibitor or TAM should be given as first line endocrine therapy [3-9].

In this Clinical Translational Research award, we will build classifiers that accurately separate antiestrogen sensitive from antiestrogen resistant breast tumors and begin to assist in the direction of specific endocrine treatments (antiestrogen *vs.* aromatase inhibitor) to *individual* patients. We hypothesize that endocrine responsiveness is affected by a gene network, rather than the activity of only one or two genes or signaling pathways [10-12]. Since the key components of such a network are unknown, we must study 10,000s of genes. We will use Affymetrix GeneChips. We will not identify mutational events, the presence of mRNA splice variants, or post-translational protein modifications. However, these factors have major effects on the transcriptome and their "footprints" should be identified by expression microarrays.

BODY

Overview: We will build classifiers that separate antiestrogen sensitive from antiestrogen resistant breast tumors and begin to assist in the direction of specific endocrine treatments (antiestrogen *vs.* aromatase inhibitor) to *individual* patients. To achieve this goal, and consistent with a CTR award, we will complete a 4-year, prospective, neoadjuvant study with Letrozole (LET) or TAM as the only systemic therapy. We will obtain molecular profiles from Affymetrix GeneChips and further develop and apply our innovative bioinformatic and biostatistic methods to explore these high dimensional data sets and build/validate new classifiers. A more accurate predictor of endocrine responsiveness would have widespread clinical use, allowing women and physicians to make more individualized and appropriate treatment decisions. For example, patients with tumors predicted to be resistant to antiestrogens and/or aromatase inhibitors would be strong candidates for an early intervention with cytotoxic chemotherapy.

In most predictive/prognostic marker studies investigators focus on a *single* factor and whether they obtain a p-value that reaches conventional statistical significance. Our approach is different because we will determine whether we can find joint gene subsets that can separate patients into sufficiently distinct groups that should differ in their treatment. We will **(1)** analyze >33,000 genes on retrospective and prospective material, **(2)** apply new biostatistical and bioinformatic methods to identify ~40 potentially informative "biomarkers," **(3)** build neural network and biostatistical model classifiers, **(4)** evaluate the joint discriminant power of selected genes concurrently rather than as single biomarkers, **(5)** focus on prediction for individual patients where the assessment of a p-value is less important than the classification rate of our predictors, **(6)** validate the classifiers in independent data sets, and **(7)** explore the ability of predictors to refine the targeting of *specific* endocrine therapies.

Evidence has begun to accumulate suggesting that an aromatase inhibitor might be a more effective first line endocrine therapy for some breast cancer patients than the current standard of care (Tamoxifen). These data have generated considerable interest and controversy, in part because unlike TAM, there are no long term studies with aromatase inhibitors where definitive survival data are available. Our study could provide new and

innovative insights into how to approach the more effective targeting of specific endocrine therapies to individual patients.

Specific Aims
We will complete two clinical studies and collect gene expression profiles from which to build predictors of endocrine responsiveness. Predictors will be built in Specific Aim 2 and validated in Specific Aim 3.

AIM 1: Clinical Studies - **Clinical Study-1** (retrospective) is of pretreatment, single, frozen samples where we will compare the molecular profiles of tumors that recurred on TAM with those of tumors that did not recur. Each resistant sample is matched with a TAM sensitive sample by age, stage, and duration of follow-up. We also have further, single (unmatched), frozen samples from patients already progressing on TAM. **Clinical Study-2** is a prospective study of breast tumor samples from patients treated with neoadjuvant TAM or LET.

AIM 2: We will develop and apply novel bioinformatics and biostatistics to discover gene subsets that define the molecular differences between endocrine sensitive and resistant breast tumors. These genes will be used, in combination with established predictive/prognostic factors, *e.g.*, ER, PgR, stage, to build innovative classifiers that can better predict an individual tumor's endocrine responsiveness.

AIM 3: We will test, optimize, and validate the performance of the classifiers from Aim 2 in retrospective studies of human breast tumors. We will measure each gene individually by IHC, *in situ* RNA hybridization (ISH), or *real time* PCR (RT-PCR).

KEY RESEARCH ACCOMPLISHMENTS

As noted in previous reports, progress on the clinical goals for this award was greatly delayed because of the time taken to obtain DOD approval of our preexisting institutionally approved IRBs at Georgetown University and at the University of Edinburgh. All institutionally approved protocols and requested material were submitted to the DOD in July 2004; additional information was requested by the DOD several months later and submitted in November 2004. We did not receive final approval to proceed with the clinical studies until March 2005. Much of this delay seems to have been entirely unavoidable (see prior reports). Clearly, this has likely left us behind schedule in recruitment to the prospective studies. We have made significant strides in our development of new analytical procedures, and we have been successful in using our emerging data to support additional applications for funding. Publications supported since the commencement of this award are listed under "Reportable Outcomes"; these constitute some of our major accomplishments in the past year. These and other key research accomplishments are presented below.

Progress on our Statement of Work (SOW from the original application)

- **TASK 1.** Array breast tumor samples from Clinical Studies 1 (retrospective) and 2 (prospective)

We have now received a total of 481 breast specimens from our collaborators at the University of Edinburgh. These specimens arrived at different times and were initially banked so that they could be processed in the most effective and logical manner. These specimens represent a mix of the initial prospective and most of the retrospective specimens. We have processed all of these specimens as frozen sections and all 481 have now been fully analyzed and annotated by the study pathologist. We have successfully extracted total RNA from 357 specimens, and labeled 169 for analysis. We have also completed the hybridization and assessment of microarray data quality control on 102 breast cancer specimens.

We requested that the specimens be sent independent of the clinical information, so that we could adequately and appropriately randomize the RNA preparation, labeling and hybridization and minimize any

operator-induced or technology-induced bias. All specimens were processed using our standard operating procedures; each manipulation being performed by the same individual to further reduced inter-operator variability. Details of the methods, quality control measures and general experimental approaches have been described in detail in earlier annual reports.

In general, we have found the material from our collaborators to be of high quality. Over 80% of the specimens have yielded RNA with an estimated RIN value of ≥5.0, which we have found produces good quality expression data. Some of the specimens with slightly lower RIN values also will be useful but we will first label, hybridize and array the best quality specimens. While not all cases generate adequate RNA, and a small number of tissues have very little neoplastic tissue (most likely the best responders), thee data suggest that we should be able to address our central hypotheses as initially anticipated.

We have also found these data to be particularly useful in supporting other studies that are ongoing in the laboratory. For example, these data have been used to support two pending R01 applications on genes we identified and described in the preliminary data for this application. We have also used these data to provide preliminary data on gene expression values that have led to our colleagues initiating other studies directed at developing therapeutic strategies to target individual genes we have identified from within this data set or from other sources.

- **TASK 2.** Store, process, and train/optimize classifiers from gene expression microarray data (modified to reflect our adoption of caArray)

As noted in our previous reports, we continue to make significant progress on addressing this task, largely as a consequence of our involvement in the National Cancer Institute Center for Bioinformatics (NCICB) led caBIG project. The PI (Dr. Clarke) leads the Lombardi Comprehensive Cancer Center's caBIG team and we have been actively involved in the development of caArray (NCICB's grid-enabled, MIAME compliant, microarray database). Indeed, we will host a key caBIG face-to-face meeting at our institute that will represent a joint meeting between the Architecture and Vocabulary and Common Data Elements Workspaces members.

With respect to the further development and optimization of data analysis algorithms, we have begun to develop novel approaches for network analysis that will allow us to further identify potentially novel targets for therapeutic intervention in subsequent studies. We anticipated obtaining such information in our original application, and we have found approaching this goal to be realistic in a much shorter time frame than initially expected.

We also continue to improve our existing algorithms and have recently submitted for publication a short communication on the implementation and uses of our VISDA algorithms (described in the initial application). As described in last year's report, we see the potential to obtain novel mechanistic insights as a significant advantage to our ongoing studies. We will provide additional information in this regard in subsequent reports; relevant publications in this area are included below in the section "Reportable Outcomes."

- **TASK 3.** Retrain/reoptimize classifiers using IHC data from Series 1 (Archival Tissues) and Series 2 (Scottish Adjuvant TAM Trial) for Validation

To perform this task we will obtain clinical information and breast tumor samples from University of Edinburgh (formalin fixed/paraffin embedded). We will rank and prioritize selected joint genes from RNA classifier built and optimized in TASK 2 (above) and retrain/reoptimize the initial neural network IHC classifier (MLP). Finally, we will validate IHC classifier on independent data sets (data sets not used to build and train the MLP classifiers).

We remain unable to move this task substantially forward on the timeframe as initially proposed because of the

delays in getting approval to work with the clinical specimens. It is addressing this aim in detail that will be the greatest consideration in assessing the need to take the reviewer's advice from their assessment of last year's report that a request for a no-cost extension may be advisable (see also "Conclusions" below).

REPORTABLE OUTCOMES

Papers and Meeting Reports*

Updates (cited as "in press" in the last report and now in print)

- Ressom, H.W., Zhang, Y., Xuan, J., Wang, Y. **& Clarke, R.** "Inference of gene regulatory networks from time course gene expression data using neural networks and swarm intelligence." *IEEE Symp Compl Intel Bioinf Comput Biol*, 435-442, 2006.

New Publications (for the present reporting period)

- Wang, L.H., Yang, X.Y., Zhang, X., An, P., Kim, H.-J., Huang, J., **Clarke, R.,** Osborne, C.K., Inman, J.K., Appella, E. & Farrar, W.L. "Disruption of estrogen receptor DNA-binding domain and related intramolecular communication restores tamoxifen sensitivity in resistant breast cancer." *Cancer Cell*, 10: 487-499, 2006.

- Bouker, K.B., Skaar, T.C., Harburger, D.S., Riggins, R.B., Fernandez, D.R., Zwart, A. **& Clarke, R.** "The A4396G polymorphism in interferon regulatory factor-1 is frequently expressed in breast cancer." *Cancer Genet Cytogenet*, 175: 61-4, 2007.

- Wong, L.-J.C., Dai, P., Lu, J.-F., Lou, M.A., **Clarke, R.** & Nazarov, N. "AIB1 gene amplification and the instability of polyQ encoding sequence in breast cancer cell lines." *BMC Cancer* 6: 111-132, 2006.

- Kuske, B., Naughton, C., Moore, K., MacLeod, K.G., Miller, W.R., **Clarke, R.** Langdon, S.P. & Cameron, D.A. "Endocrine therapy resistance can be associated with high estrogen receptor alpha (ERα) expression and reduced ERα phosphorylation in breast cancer models." *Endocr Related Cancer*, 13: 1121-1133, 2006.

- Gong, T., Xuan, J., Zhu, J., Li, H., **Clarke, R.,** Hoffman, E. & Wang, Y. "Composite gene module discovery using non-negative independent component analysis." *IEEE/NLM Life Sci Sys Apps Workshop*, 1-3, 2006.

- Feng, Y., Wang, Z., Zhu, Y., Xuan, J., Miller, D., **Clarke, R.,** Hoffman, E.P. & Wang, Y. "Learning the tree of phenotypes using genomic data and VISDA." *6th IEEE Symp Bioinf Bioeng (BIBE '06)*, 165-170, 2006

- Gong, T., Zhu, Y., Xuan, J., Li, H., **Clarke, R.,** Hoffman, E.P. & Wang, Y. "Latent variable and nICA modeling of pathway gene module composite." *Proc 28th IEEE EMBS Intl Conf*, pp. 5872-5875, 2006.

- Ressom, H.W., Zhang, Y., Xuan, J., Wang, Y. & **Clarke, R.** "Inferring network interactions using recurrent neural networks and swarm intelligence," *Proc 28th IEEE EMBS Intl Conf*, pp. 4241-4244, 2006.

*We include in the appendix reprints of those papers that are already published adn for which we have proofs or reprints. We do not list here or include in the appendices any published abstracts, but can do so if requested.

Several other manuscripts related to our bioinformatic methods also are submitted and in preparation – these will be cited reported in the next report. Please note that the papers published in the engineering literature are different from most conference proceedings in the biomedical literature. These are not abstracts but fully peer-reviewed publications comparable to short communications in biomedical journals.

Comment on Subcontracts: Please also note that the majority of our publications here and in prior years include coauthors from one or both of our subcontracts. Thus, our program is working very effectively and collaboratively, this should further be apparent in the development of new informatics methods (Virginia Polytechnic and State University subcontract) and the large number of high quality breast tumor specimens we have obtained from the University of Edinburgh.

CONCLUSIONS

We have made good progress on the research infrastructure goals and in the development or optimization of the methods needed for data analysis. We also have completed and published all of the data presented as preliminary data in the initial application. The clinical studies were held up by an unexpectedly long delay in obtaining final approval for our existing protocols. As noted by the reviewer of last year's annual report, this delay will adversely affect the prospective study and this reviewer indicated that a request for a one-year no cost extension might be required. We concur that this may yet be required and will evaluate the need for this over the next six months. If deemed necessary, we will apply for such an extension in writing before the end of the original funding period, this should ensure that we remain in compliance with USAMRMC guidelines, maintain continuity of the project and successfully complete our studies.

REFERENCES

1. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomized trials. Lancet, *351:* 1451-1467, 1998.

2. Early Breast Cancer Trialists Collaborative Group: Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. Lancet, *399:* 1-15, 1992.

3. Winer, E. P., Hudis, C., Burstein, H. J., Chlebowski, R. T., Ingle, J. N., Edge, S. B., Mamounas, E. P., Gralow, J., Goldstein, L. J., Pritchard, K. I., Braun, S., Cobleigh, M. A., Langer, A. S., Perotti, J., Powles, T. J., Whelan, T. J., and Browman, G. P. American Society of Clinical Oncology technology assessment on the use of aromatase inhibitors as adjuvant therapy for women with hormone receptor-positive breast cancer: status report 2002. J Clin Oncol, *20:* 3317-3327, 2002.

4. Ravdin, P. Aromatase inhibitors for the endocrine adjuvant treatment of breast cancer. Lancet, *359:* 2126-2127, 2002.

5. Bonneterre, J., Buzdar, A., Nabholtz, J. M., Robertson, J. F., Thurlimann, B., von Euler, M., Sahmoud, T., Webster, A., and Steinberg, M. Anastrozole is superior to tamoxifen as first-line therapy in hormone receptor positive advanced breast carcinoma. Cancer, *92:* 2247-2258, 2001.

6. Baum, M., Buzdar, A. U., Cuzick, J., Forbes, J., Houghton, J. H., Klijn, J. G., Sahmoud, T., and ATAC Trialists Group Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. Lancet, *359:* 2131-2139, 2002.

7. Ellis, M. J., Coop, A., Singh, B., Mauriac, L., Llombert-Cussac, A., Janicke, F., Miller, W. R., Evans, D. B., Dugan, M., Brady, C., Quebe-Fehling, E., and Borgs, M. Letrozole is more effective neoadjuvant

endocrine therapy than tamoxifen for ErbB-1- and/or ErbB-2-positive, estrogen receptor- positive primary breast cancer: evidence from a phase III randomized trial. J Clin Oncol, *19:* 3808-3816, 2001.

8. Miller, W. R., Anderson, T. J., and Dixon, J. M. Anti-tumor effects of letrozole. Cancer Invest, *20 Suppl 2:* 15-21, 2002.

9. Smith, I. E. and Dowsett, M. Aromatase inhibitors in breast cancer. N Engl J Med, *348:* 2431-2442, 2003.

10. Clarke, R., Leonessa, F., Welch, J. N., and Skaar, T. C. Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol Rev, *53:* 25-71, 2001.

11. Clarke, R. and Brünner, N. Acquired estrogen independence and antiestrogen resistance in breast cancer: estrogen receptor-driven phenotypes? Trends Endocrinol Metab, *7:* 25-35, 1996.

12. Clarke, R., Skaar, T. C., Bouker, K. B., Davis, N., Lee, Y. R., Welch, J. N., and Leonessa, F. Molecular and pharmacological aspects of antiestrogen resistance. J Steroid Biochem Mol Biol, *76:* 71-84, 2001.

IEEE Symp Comp Intel Bioinformatics Comp Biol 435-442, 2006

Inference of Gene Regulatory Networks from Time Course Gene Expression Data Using Neural Networks and Swarm Intelligence

H. W. Ressom[1], Y. Zhang[1,2], J. Xuan[2], Y. Wang[2], R. Clarke[1]

[1]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC USA
[2]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA USA

Abstract— **We present a novel algorithm that combines a recurrent neural network (RNN) and two swarm intelligence (SI) methods to infer a gene regulatory network (GRN) from time course gene expression data. The algorithm uses ant colony optimization (ACO) to identify the optimal architecture of an RNN, while the weights of the RNN are optimized using particle swarm optimization (PSO). Our goal is to construct an RNN whose response mimics gene expression data generated by time course DNA microarray experiments. We observed promising results in applying the proposed hybrid SI-RNN algorithm to infer networks of interaction from simulated and real-world gene expression data.**

I. INTRODUCTION

Gene regulatory network (GRN) is a model of a network that describes the relationships among genes in a given condition. The model can be used to enhance the understanding of gene interactions and better ways of elucidating environmental and drug-induced effects.

Large-scale monitoring of gene expression such as DNA microarrays [1-4] is considered to be one of the most promising techniques for making the discovery of GRNs feasible [5]. However, the task of inferring GRNs involves several challenges including the following: (1) the number of related genes is very large compared to the number of samples or time points, (2) the observed data involve a significant amount of noise, and (3) the interaction among genes displays complex (nonlinear and dynamic) relationships. This challenge provides computer scientists, statisticians, and engineers with opportunities to expand their knowledge of intelligent methods to provide models for better understanding of biological systems.

The field of system modeling plays a significant role in the discovery of GRNs. Several system modeling approaches have been proposed to reverse-engineer network interactions including a variety of continuous or discrete, static or dynamic, quantitative or qualitative methods [1-6].

The use of computational intelligence (CI) methods for system modeling has gained particular interest, because they require little a priori knowledge about the underlying system and the model can be derived from data. Given a set of data, these methods discover hidden regularities and structures within the data. Instead of requiring patterns to be known ahead of time, they search automatically for patterns that are hidden in the data.

Several CI methods have been proposed [7-14] for GRN reconstruction. These methods have been successfully applied in both cluster- and classification-based approaches. For example, in [9], a neural model is used to simulate the dynamics of the lambda phage regulatory system. Middendorf et al. [15] used decision trees to predict whether a gene is up- or down-regulated in a particular experiment on the basis of the presence of binding site subsequences (motifs) in the gene's regulatory region and the expression level of regulators such as transcription factors in the experiment. Soinov et al. [16] applied decision-tree based classifier to extract simple rules defining gene interrelations. In [17], an approach is proposed based on fuzzy rules of a known activator/repressor model of gene interaction. This algorithm transforms expression values into qualitative descriptors that can be evaluated by using a set of heuristic rules and searches for regulatory triplets consisting of activator, repressor, and target gene. This approach, though logical, is a brute force technique for finding gene relationships. It involves a significant computation time, which restricts its practical usefulness. Also, this method is limited to the study of the interaction between one possible positive and one negative regulator for each gene. In [18], we proposed the use of clustering as an interface to a fuzzy logic-based method to improve the computational efficiency. A scalable linear variant of fuzzy logic is introduced in [19] to examine the interactions of multiple genes. Genetic algorithms (GAs) have also been applied to decipher genetic networks from gene expression data [10-12]. Shin and Iba [12] developed an inference algorithm based on GAs for the optimization of the influence matrix of GRN. In [14], GAs and ANNs are combined to determine gene interactions in temporal gene expression data.

To capture the nonlinear and dynamic relationships, we propose to model GRNs using recurrent neural networks (RNNs), which consist of nonlinear processing elements (neurons) that possess feedback and memory units. The

architecture and the synaptic weights of RNNs are optimized using two recently introduced swarm intelligence (SI) methods, ant colony optimization (ACO) and particle swarm optimization (PSO) methods, respectively. The hybrid SI-RNN algorithm is applied to infer networks of interactions from simulated and real-world gene expression data, yielding promising results.

The paper is organized as follows. Section II highlights the steps involved in using CI for system modeling. Section III describes our proposed SI-RNN algorithm for inferring network interactions. Section IV presents networks inferred from simulated and yeast cell cycle gene expression data. Finally, Section V concludes the paper.

II. Systems Modeling using CI Methods

Besides selecting a suitable CI paradigm, developing an intelligent model involves four steps: data preparation, model structure selection, learning, and model evaluation. These steps are repeated until the last step results in a satisfactory performance.

Many times the "raw" data are not the best data to use for modeling a CI paradigm. Hence, in using CI paradigms to solve real-world problems, it is important to transform raw data into a form acceptable to the paradigm. The first step is to decide what the inputs and outputs are. Inputs that are not relevant for modeling should be excluded. The next step is to process the data in order to handle missing data, remove outliers, and to normalize and scale the data into acceptable range. This will be followed by model structure selection, which includes the choice of neural network architecture, fuzzy rules, membership functions, fuzzy operators, genetic operators, and coding scheme. The selection of neural network architecture includes choosing activation functions, appropriate number of layers, number of neurons in each layer, and the interconnection of the neurons and the layers. After the model structure is selected, its free parameters will be determined. The predominant feature of CI paradigms is that they learn from data. We define learning as the process of finding the free parameters of a CI model (e.g. determining the weights of a neural network). After learning is completed, a CI paradigm is evaluated for its performance through testing. The purpose of this testing is to prove the adequacy or to detect the inadequacy of the fitted model. The latter could arise from an inappropriate selection of network topology, too small or too many neurons, or from insufficient training or overtraining. Incorrect input node assignments, noisy data, error in the program code, or several other effects may also cause a poor fit. The aim of model evaluation is to insure that the model fit is correct; that the model satisfies the desired requirements, and that it serves as a general model. A general model is one whose input-output relationships (derived from the training dataset) apply equally well to new sets of data (previously unseen test data) from the same problem not included in the training set. The main goal of intelligent modeling is thus the generalization to new data of the relationships learned on the training set.

The choice of appropriate CI paradigm is critical to the modeling process. This primarily depends on the complexity of the underlying system to be modeled, the available information (a priori knowledge and data), and the existence of a suitable learning algorithm. In recent years, ANNs have been employed successfully for modeling a wide range of nonlinear systems. They generally consist of a number of interconnected processing elements known as neurons. The way neurons are interconnected or how the inter-neuron connections are arranged determines the architecture of a neural network. The strengths of the connections (known as weights or synaptic weights) are adjusted or trained to achieve a desired overall behavior of the network. The most popular architecture is a feedforward neural network, where the neurons are grouped into layers. All connections are feedforward; that is, they allow information transfer only from an earlier layer to the next consecutive layers. It is known that ANN can sufficiently approximate the nonlinear mapping, learn to adapt to dynamics of uncertain systems, and have strong robustness and fault-tolerant abilities due to the rich connection and nonlinear activation functions of the neurons. In light of the above advantages, neural network based approaches have shown the superiority over the well-established and proven conventional methods for parameter/state estimation for a large class of problems and systems. However, in order to perform a time series prediction or build a model of a dynamical system such as one that represents regulatory interactions, it is important to establish a form of expansion to the feedforward neural network, so that the network contains some type of memory element. This can be achieved by applying time-delayed inputs to feed forward networks. Alternatively, an RNN can be built, in which the outputs of some neurons are fed back to the same neurons or to other neurons in the network. Thus, signals can flow in both forward and backward directions. RNNs have a dynamic memory - their outputs at a given instant reflect the current model input as well as previous inputs and outputs. It is an ideal candidate for modeling dynamic systems such as network interactions in biological systems.

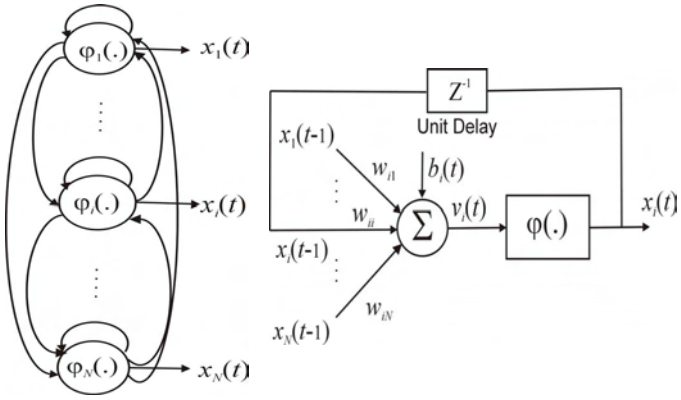
III. Inferring Network Interaction Using SI-RNN

In building an RNN to infer a network of interactions, the identification of the correct structure and determination of the free parameters (weights and biases) to mimic measured data is a challenging task given the limited available quantity of data. For example, in inferring a GRN from microarray data, the number of time points is considerably low compared to the number of genes involved. Considering the complexity of the biological system, it is difficult to adequately describe the pathways involving a large number of genes with few time points. In this paper, we apply ACO and PSO methods to select the optimal architecture of an RNN and to update its free parameters, respectively. ACO is a discrete optimization algorithm that has been successfully used for combinatorial

problems, while PSO is applied for continuous optimization (i.e., the variables in the objective function can assume real values). We formulated the selection of RNN structures as a combinatorial problem that can be effectively optimized by ACO. The optimal parameters of an RNN for a given structure can be obtained by using PSO as this is a continuous optimization problem.

A. Recurrent Neural Network

Figure 1a shows an RNN, where the output of each neuron is fed back to its input after a unit delay and is connected to other neurons. It can be used as a model of gene regulatory network, where every gene in the network is considered as a neuron. The RNN can model not only the interactions between genes but also gene self-regulation, which is represented by a unit delay (z^{-1}).



(a) (b)

Figure 1. (a) Architecture of a fully connected RNN; (b) Details of a single recurrent neuron.

Figure 1b illustrates the details of the *i*th self-feedback neuron (e.g. *i*th gene in the GRN), where v_i, known as the induced local field (activation level), is the sum of the weighted inputs (the regulation of other genes) to the neuron (*i*th gene); and φ(.) represents an activation function (integrated regulation of the whole RNN on *i*th gene), which transforms the activation level of a neuron into an output signal (regulation result). The induced local field and the output of the neuron, respectively, are given by:

$$v_i(t) = \sum_{j=1}^{N} w_{ij} x_j(t-1) + b_i \qquad (1)$$

$$x_i(t) = \varphi(v_i(t)) \qquad (2)$$

where the synaptic weights w_{i1}, w_{i2},…, w_{iN} define the strength of connection between the *i*th neuron (e.g. *i*th gene) and its inputs (e.g. expression level of genes). Such synaptic weights exist between all pairs of neurons in the network. b_i denotes the bias for the *i*th neuron. We denote $\vec{s}$ as a structure vector that describes the architecture of the network, and $\vec{w}$ as a weight vector that consists of all the synaptic weights and biases in the network. $\vec{s}$ and $\vec{w}$ are adapted during learning to yield the desired network outputs. The activation function φ introduces nonlinearity to the model. When information about the complexity of the underlying system is available, a suitable activation function can be chosen (e.g. linear, logistic, sigmoid, threshold, hyperbolic tangent sigmoid or Gaussian function.) If no prior information is available, our algorithm uses the hyperbolic tangent sigmoid function.

As a cost function, we use the mean-squared error between the expected output and the network output across time (from the initial time point t_0 to the final time point t_f) and across neurons in the network. The cost function can be written as:

$$E(\vec{w}) = \frac{1}{t_f N} \sum_{t=t_0}^{t_f} \sum_{i=1}^{N} [x_i(t) - \hat{x}_i(t)]^2 \qquad (3)$$

where $x_i(t)$ and $\hat{x}_i(t)$ are the true and predicted values (expression levels) for the *i*th neuron (gene) at time *t*. The goal is to determine the structure vector $\vec{s}$ and weight vector $\vec{w}$ that minimize this cost function. We propose ACO and PSO to optimize $\vec{s}$ and $\vec{w}$, respectively.

Note that the above RNN is self-evolutionary and can be used to model a multi-step-ahead prediction. The RNN starts with a given initial condition, evolves, and eventually reaches final states. In this study, we inferred a network from a simulated dataset generated by a five-node network using both one-step-ahead and multi-step-ahead forms, which resulted in similar performance. For real-world gene expression data that involve noise and unequal time intervals, we use the one step-ahead prediction method during training. However, the resulting RNN can be used to simulate multiple-ahead predictions.

B. Ant Colony Optimization

Ant colony optimization studies artificial systems that take inspiration from the behavior of real ant colonies. The basic idea of ACO is that a large number of simple artificial agents are able to build good solutions to solve hard combinatorial optimization problems via low-level based communications.

We propose to use ACO to optimize the structure vector $\vec{s}$. Each possible network structure $\vec{s}$ is defined by a combination of *n* features $\vec{s} = [s_1\ s_2 \cdots s_n]$, where s_j is an *n*-bit binary string that indicates which neurons are controlled by neuron *j*. Each s_j is selected from 2^n candidate features. For each neuron *j*, we define the function in Eq. (4) to determine the probability of selecting a feature *i* among the 2^n candidate features:

$$P_i^j(k) = \frac{\tau_i^j(k)}{\sum_{i=1}^{n} \tau_i^j(k)} \qquad j = 1,..,n \qquad i = 1,...,2^n \qquad (4)$$

where $\tau_i^j(k)$ is the amount of pheromone trail for the *i*th feature at iteration *k*. At *k*=0, $\tau_i^j(k)$ is set to a constant for all features, allowing each feature to have equal probability of being selected. Thus, in the first iteration, each ant chooses randomly *n* features that make up a structure ($\vec{s}$, a trail). Let

$\vec{s}$ be an ant consisting of n features $\vec{s} = [s_1\ s_2 \cdots s_n]$. Depending on the performance of $\vec{s}$, the amount of pheromone trail of all features in $\vec{s}$ is updated. The performance function here is evaluated on the basis of mimicking the response of the system under study. To estimate the performance of $\vec{s}$, we construct an RNN that has the structure defined in $\vec{s}$. Then, we optimize the weights of the RNN using PSO. The response of the resulting RNN will be compared with the measured/observed response of the system under study. The amount of pheromone trail for each element in $\vec{s}$ is updated in proportion to the performance of the structure. Assuming that the *i*th feature for the *j*th neuron was in $\vec{s}$, the corresponding amount of pheromone trail will be updated as follows:

$$\tau_i^j(k+1) = \rho.\tau_i^j(k) + \Delta\tau_i(k) \quad (5)$$

where ρ is a constant between 0 and 1, representing the evaporation of pheromone trails, and $\Delta\tau_i(k)$ is an amount proportional to the performance by $\vec{s}$. $\Delta\tau_i(k)$ is set to zero, if $s_i \notin \vec{s}$. This update is made for all N ants ($\vec{s}_1, \ldots, \vec{s}_N$). Note that at k=0, $\Delta\tau_i(k)$ is set zero for all features. The updating rule allows trails that yield good performance to have their amount of pheromone trail increased, while others will evaporate. As the algorithm progresses, features with large amounts of pheromone trails influence the probability function to lead the ants towards them.

C. Particle Swarm Optimization

In the PSO algorithm, each particle is represented as a vector $\vec{w}_i$ and instantaneous trajectory vector $\Delta\vec{w}_i(k)$, describing its direction of motion in the search space at iteration k. The index i refers to the ith particle. The core of the PSO algorithm is the position update rule (6) which governs the movement of each of the n particles through the search space.

$$\vec{w}_i(k+1) = \vec{w}_i(k) + \Delta\vec{w}_i(k+1)$$
$$\Delta\vec{w}_i(k+1) = \chi(\Delta\vec{w}_i(k) + \Phi_1(\vec{w}_{i,best}(k) - \vec{w}_i(k)) + \Phi_2(\vec{w}_{G,best}(k) - \vec{w}_i(k)))$$

where

$$\Phi_1 = c_1\begin{bmatrix} r_{1,1} & 0 & 0 & 0 \\ 0 & r_{1,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_{1,D} \end{bmatrix} \text{ and } \Phi_2 = c_2\begin{bmatrix} r_{2,1} & 0 & 0 & 0 \\ 0 & r_{2,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_{2,D} \end{bmatrix} \quad (6)$$

At any instant, each particle is aware of its individual best position, $\vec{w}_{i,best}(k)$, as well as the best position of the entire swarm, $\vec{w}_{G,best}(k)$. The parameters c_1 and c_2 are constants that weight particle movement in the direction of the individual best positions and global best positions, respectively; and $r_{1,j}$ and $r_{2,j}$, $j = 1,2,\ldots D$ are random scalars distributed uniformly between 0 and 1, providing the main stochastic component of the PSO algorithm.

The constriction factor, χ, may also help to ensure convergence of the PSO algorithm, and is set according to the weights c_1 and c_2 as in (7).

$$\chi = \frac{2}{\left|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}\right|}, \quad \varphi = c_1 + c_2, \quad \varphi > 4 \quad (7)$$

The key strength of the PSO algorithm is the interaction among particles. The second term in (6), $\Phi_2(\vec{w}_{G,best}(k) - \vec{w}_i(k))$, is considered to be a "social influence" term. While this term tends to pull the particle towards the globally best solution, the first term, $\Phi_1(\vec{w}_{i,best}(k) - \vec{w}_i(k))$, allows each particle to think for itself. The net combination is an algorithm with excellent trade-off between total swarm convergence, and each particle's capability for global exploration. Moreover, the relative contribution of the two terms is weighted stochastically.

The algorithm consists of repeated application of the velocity and position update rules presented above. Termination can occur by specification of a minimum error criterion, maximum number of iterations, or alternately when the position change of each particle is sufficiently small as to assume that each particle has converged.

Selection of appropriate values for the free parameters of PSO plays an important role in the algorithm's performance. In our study, parameters c_1 and c_2 were arbitrarily selected (c_1 = 2.05, c_2 = 2.05), with constriction factor, χ, determined by (7) and the maximum velocity was set at 2.

D. SI-RNN

In this section, we illustrate how the two SI methods, ACO and PSO, work together to optimize both $\vec{s}$ and $\vec{w}$ of an RNN to mimic the response of an unknown network of interactions. Each node in the true network is represented by a neuron in the RNN. We assume that the number of nodes in the network is known (e.g. the number of genes for which a gene regulatory network is to be modeled is known), but the way the nodes interact is assumed to be unknown.

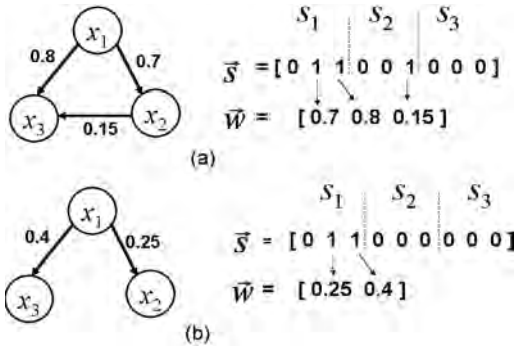


Figure 2. True network (a); randomly selected network (b).

ACO starts with initial candidate ants that define various structures. For example, let Fig. 2(a) be the structure for the true interaction that has three nodes and Fig. 2(b) be one of the randomly selected initial structures by ACO. For each of these

two systems, the corresponding structure vector $\vec{s}$ is shown in the figure, where s_j (j = 1, 2, 3) is a three-bit binary string that indicates which neurons (including itself) are controlled by the ith neuron. For example, if s_1 = [0 1 1], it implies that the first neuron controls all others except itself and s_2 = [0 0 1] implies that the second neuron controls the third neuron only.

PSO searches for the optimal weight vector $\vec{w}$ to minimize the difference between the output of the true network and the RNN using the training data. Only the elements of $\vec{w}$ that correspond to nonzero entries in $\vec{s}$ are updated by PSO. For example, the weight vector $\vec{w}$ in Fig. 2b contains only two variables: 0.25 and 0.4, corresponding to the two nonzero entries in the structure vector $\vec{s}$.

The optimal weight vectors for all randomly selected structures (ants) are tested with previously unseen validation data. The performance of each particle in simulating the validation data is returned to ACO to update the trails of the ants in the search space (i.e., update the structure vector $\vec{s}$). The new s_j's in $\vec{s}$ are used to construct a new candidate structure. This will lead to a structure that is more similar to the global best structure than the previous one. The assumption in this algorithm is that the prediction error of an arbitrary network will be larger than a network that matches the correct structure. Through subsequent iterations, the ACO and PSO search for the optimal structure vector $\vec{s}$ and weight vector $\vec{w}$ to make accurate predictions.

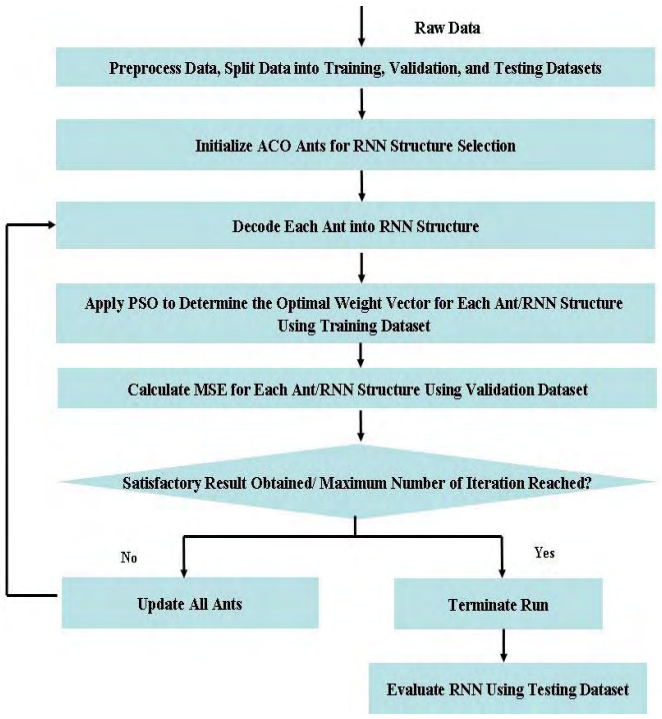


Figure 3. Flowchart for the SI-RNN algorithm.

Figure 3 illustrates the overall algorithm, which involves the two major components: (i) ACO that initially generates various topologies randomly and selects the most optimal structure iteratively based on the prediction performance of the topologies in a validation dataset and (ii) PSO that determines the free parameters of a given topology with the aim of minimizing the error between the training data and the outputs of the networks whose topologies are determined by ACO. The performance of the PSO particles in predicting the validation dataset is used to iteratively update the network structure. The process continues until satisfactory result is found or maximum number of iterations is reached. The final network will be evaluated via independent testing dataset.

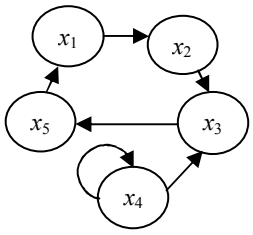
IV. Data and Results

The SI-RNN algorithm is evaluated in inferring networks of interactions from artificial and yeast cell cycle data.

A. *Artificial Data*

We applied the SI-RNN approach to identify the network in Fig. 4. We generated three datasets (training, validation, and testing) with different initial conditions. Each dataset consisted of 20 time points. These artificial data sets are created to ascertain the ability of the algorithm to "rediscover" the underlying network that generated the data.

An RNN model of five neurons with hypothetic tangent sigmoid activation function was trained using the SI-RNN algorithm. PSO used the training dataset to determine the optimal weight vector $\vec{w}$ for each structure vector $\vec{s}$ defined by ACO. The performance of each structure in predicting the outputs of the network in the validation dataset is used by ACO to determine the optimal structure. The algorithm was run 100 times. In each run, Eq. 3 was evaluated 1000 times to identify the structure that leads to the least cost. 54 runs (out of 100) predicted a RNN with identical structure to Fig. 4. 16% of the runs also predicted the true network structure, but they had two or three additional connections. The remaining 30% consisted of arbitrary structures. Fig. 5 shows the outputs of the true network and the predicted RNN for the testing dataset.



$$x_1(t) = 0.5x_5(t-1)$$
$$x_2(t) = 0.3x_1(t-1)$$
$$x_3(t) = -0.6x_2(t-1) + x_4(t-1)$$
$$x_4(t) = 0.7(x_4(t-1))^2$$
$$x_5(t) = -0.8\sqrt{|x_3(t-1)|}$$

Figure 4. A simulated five-node network.

B. *Yeast Cell Cycle Data*

The yeast cell cycle data collected by Spellman *et al.* [20] consist of six time series (cln3, clb2, alpha, cdc15, cdc28 and elu) expression measurements of the transcript (mRNA) levels of *S. cerevisiae* genes.

To test our approach, we chose five cyclin genes (CLB1, CLB2, CLB5, CLB 6 and CDC28), which are involved in cell-cycle regulation. The expression levels of these genes in three time series measurements were used to construct a GRN. In

these datasets, the biological samples were synchronized by three different methods: α factor arrest, arrest of a cdc15, and cdc28 temperature-sensitive mutant. We used the cdc15 dataset, which has 24 experimental conditions, as training dataset. The other two datasets, alpha and cdc 28, which have 18 and 17 time points, were used as validation and testing datasets, respectively.

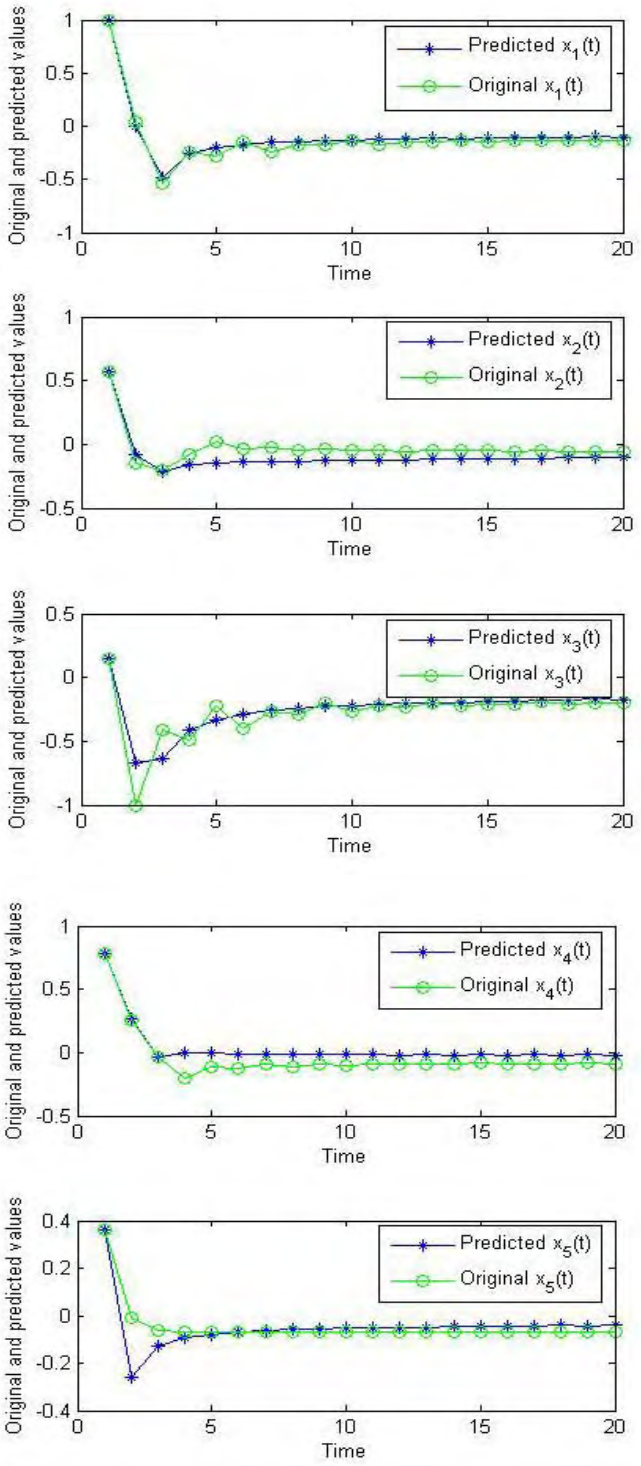


Figure 5. Original and predicted outputs of the testing dataset.

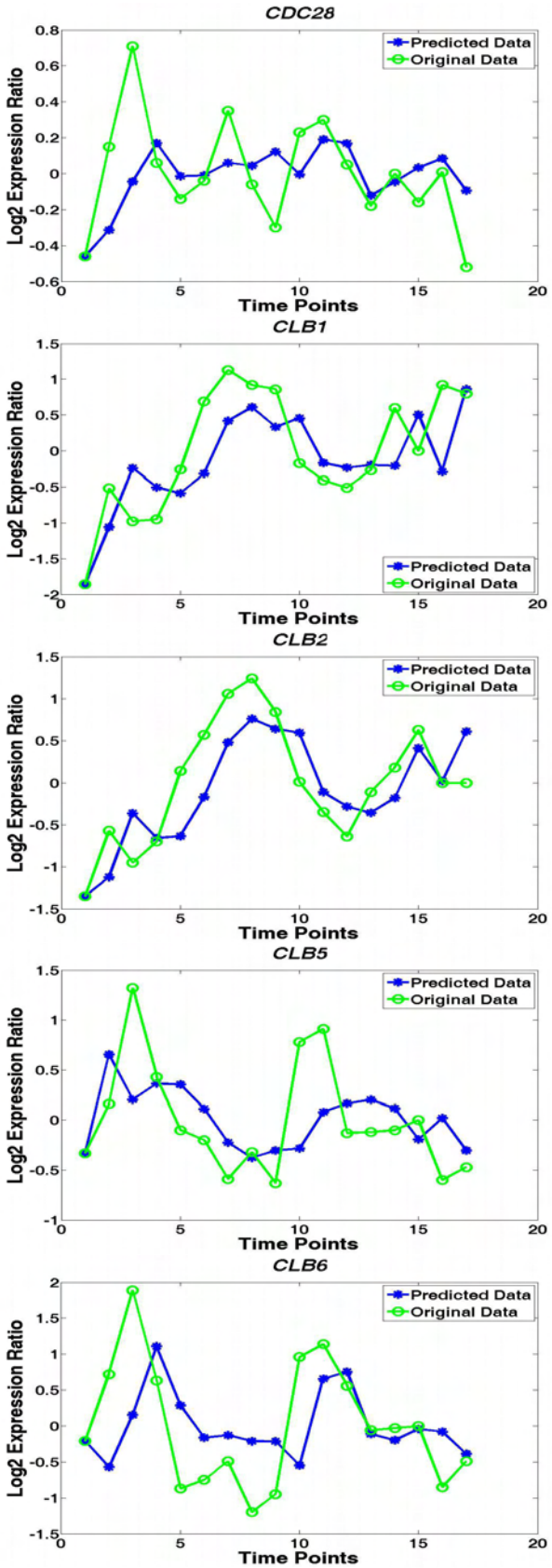


Figure 6. Original and predicted outputs of five genes in the testing dataset (cdc28 dataset).

PSO used the training dataset to determine the optimal weight vector $\vec{w}$ for each structure vector $\vec{s}$ defined by ACO. The performance of each structure in predicting the outputs of the network in the validation dataset is used by ACO to determine the optimal structure. The algorithm was run 10 times. In each run, Eq. 3 was evaluated 1000 times to identify the structure that leads to the least MSE. To improve the prediction accuracy, only the connections obtained in at least 50% of the runs were selected. Fig. 6 shows the measured gene expression data and the outputs of the predicted RNN for each gene in the testing dataset.

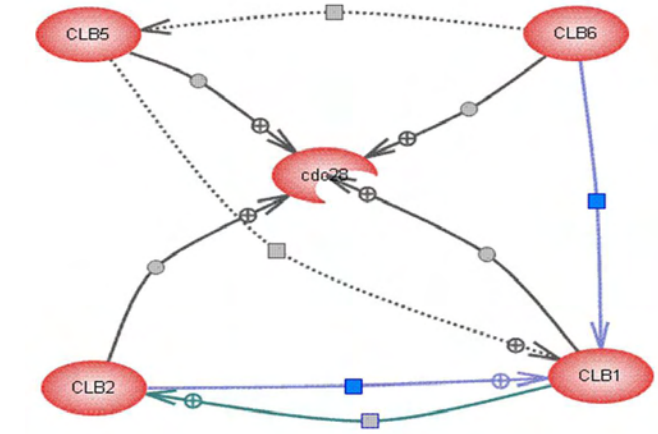
As the true GRN that governs the interaction among the five genes is not available, the accuracy of the network is determined by how well it fits the measured gene expression data. To get insight into the performance of the SI-RNN method, we randomly generated 100 RNN structures and optimized their parameters using PSO. The average and standard deviation of the MSE for the 100 randomly generated RNN were 0.25 and 0.14, respectively, while the optimal RNN structure found through the SI-RNN method yielded an average MSE = 0.1 and standard deviation = 0.08 in 10 runs. Note that the MSE is calculated for the data normalized between 0 and 1.

To further evaluate the GRN identified by SI-RNN, we searched for known interactions in a *S. cerevisiae* database provided by the PathwayStudio software (Ariadne Genomics, Rockville, MD). Figures 7a and 7b depict the interactions among the five genes found by the PathwayStudio software and by our hybrid SI-RNN algorithm, respectively. Table I illustrates the gene relationships depicted in Fig. 7.

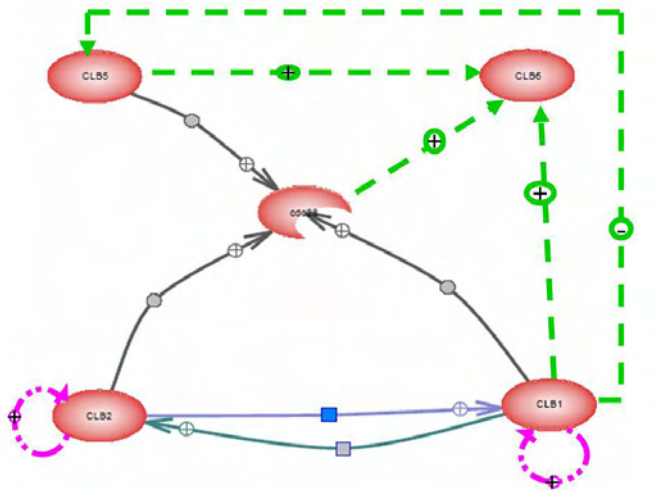
Among the 11 relations predicted by SI-RNN, five of them concur with the known interactions obtained by PathwayStudio. Two gene self-regulations (CLB1 and CLB2) were found by SI-RNN, but not by PathwayStudio. Four other relations (presented by dotted lines in Fig. 7b) are also found by PathwayStudio, but with reversed direction of regulation. PathwayStudio searches for known interactions on the basis of literature mining through natural language processing methods. From the literature used by PathwayStudio, it appears that the three "reversed" relations are described to have some relations without a defined direction of regulation. For example, CLB5 and CLB6 are essential for sporulation because they are required for premeiotic DNA replication [21]. This indicates that there maybe some relationship between these two genes, but not necessarily a directed regulation from CLB5 to CLB6. Another example is the relationship between CLB6 and CDC28. In [22], it is stated that the actual initiation events require the activities of at least two protein kinases, the cyclin-dependent kinase (CDK) Cdc28p associated with cyclin B (Clb5p or Clb6p) and the Cdc7p kinase associated with its regulatory subunit Dbf4p. No direct regulation information is provided.

TABLE I
KNOWN RELATIONS AMONG FIVE GENES FROM PATHWAYSTUDIO SOFTWARE.

Relation Type	Symbol	Predicted by SI-RNN
Expression	CLB1 <--- CLB6	yes (reversed)
Expression	CLB1 <--- cdc28	no
Expression	CLB2 <--- cdc28	no
Expression	CLB1 <+-- CLB2	yes
Regulation	CLB6 ---> CLB5	yes (reversed)
Regulation	CLB6 --+> CLB2	no
Regulation	CLB1 <+-- CLB5	yes (reversed)
MolSynthesis	CLB1 --+> CLB2	yes
Genetic Interaction	CLB2 ---- cdc28	no
Direct Regulation	CLB6 --+> cdc28	yes (reversed)
Direct Regulation	CLB5 --+> cdc28	yes
Direct Regulation	CLB2 --+> cdc28	yes
Direct Regulation	CLB2 <+-- CLB5	no
Direct Regulation	CLB1 --+> cdc28	yes



(a)



(b)

Figure 7. Cell cycle pathway in five genes: CDC28, CLB1, CLB2, CLB5 and CLB6. (a) Known interactions found by the PathwayStudio software, (b) Result of the proposed hybrid SI-RNN method: correctly identified relations use the same line as (a), reversed relations use dotted lines, and two additional self-regulations are indicated (CLB1 and CLB2).

V. Conclusion

In this paper, we explored the combined advantages of the nonlinear and dynamic properties of RNN, and the global search capabilities of swarm intelligence methods to infer network interactions. We evaluated the performance of the algorithm using data generated from a simulated five-node network, and time course gene expression data for five yeast genes. Although the algorithm yielded promising result in predicting the simulated network, due to the stochastic properties of the algorithms, not all runs identify the correct network structure. Our future work will focus on improving the rate at which the correct structure is identified.

In inferring the GRN that govern the five genes, about half of the relations predicted by SI-RNN were verified by published literature. Three additional interactions are also verified in the literature to have some relations, although no specific direction of relation was given. However, these relations predicted by SI-RNN were reversed compared to the output of the PathwayStudio software. Two self-regulations were predicted by SI-RNN, but they could not be verified by the PathwayStudio software.

Due to noise and insufficient time points in real gene expression data, we anticipate challenges in applying the proposed hybrid SI-RNN method to infer gene regulatory networks that involve large number of genes. To address these challenges, we plan to incorporate known gene interactions and genomic information into the SI-RNN algorithm.

References

[1] Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W., "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks," Bioinformatics, vol. 18, No. 2, pp. 261-74, 2002.

[2] Schmitt, W. J., Raab, R. M., and Stephanopoulos, G., "Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data," Genome Research, vol. 14, pp. 1654-1663, 2004.

[3] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., "Using Bayesian networks to analyze expression data," J Comput Biol., vol. 7, pp. 601-620, 2000.

[4] D'Haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R., "Linear modeling of mRNA expression levels during CNS development and injury," Pac Symp Biocomput, pp. 41-52, 1999.

[5] Chen, T., He, H. L., and Church, G. M., "Modeling gene expression with differential equations," Pac Symp Biocomput, pp. 29-40, 1999.

[6] Liang, S., Fuhrman, S., and Somogyi, R., "Reveal, a general reverse engineering algorithm for inference of genetic network architectures," Pac Symp Biocomput, pp. 18-29, 1998.

[7] Mendes, P. and Kell, D., "On the analysis of the inverse problem of metabolic pathways using artificial neural networks," BioSystems, vol. 38, pp. 15-28, 1996.

[8] Mendes, P., Sha, W., and Ye, K., "Artificial gene networks for objective comparison of analysis algorithms," Bioinformatics, vol. 19, pp. ii122-ii129, 2003.

[9] Vohradsky, J., "Neural model of the genetic network," J Biol Chem, vol. 276, No. 39, pp. 36168-73, 2001.

[10] Ueda, T., Ono, I., and Okamoto, M., "Development of system identification technique based on real-coded genetic algortihm," Genome Informatics, vol. 13, pp. 386-387, 2002.

[11] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M., "Dynamic modeling of genetic networks using genetic algorithm and S-system," Bioinformatics, vol. 19, No. 5, pp. 643-50, 2003.

[12] Shin, A. and Iba, H., "Construction of genetic network using evolutionary algorithm and combined fitness function," Genome Informatics vol. 14, pp. 94-103, 2003.

[13] Engelbrecht, A. P., Computational Intelligence: An Introduction: John Wiley, New York, 2003.

[14] Keedwell, E. and Narayanan, A., "Discovering gene regulatory networks with a neural-genetic hybrid," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 2, No. 3, pp. 231-243, 2005.

[15] Middendorf, M., Kundaje, A., Wiggins, C., Freund, Y., and Leslie, C., "Predicting genetic regulatory response using classification," Bioinformatics, vol. 20 Suppl 1, pp. I232-I240, 2004.

[16] Soinov, L. A., Krestyaninova, M. A., and Brazma, A., "Towards reconstruction of gene networks from expression data by supervised learning," Genome Biol, vol. 4, No. 1, pp. R6, 2003.

[17] Woolf, P. J. and Wang, Y., "A fuzzy logic approach to analyzing gene expression data," Physiol Genomics, vol. 3, No. 1, pp. 9-15, 2000.

[18] Ressom, H., Reynolds, R., and Varghese, R. S., "Increasing the efficiency of fuzzy logic-based gene expression data analysis," Physiol Genomics, vol. 13, No. 2, pp. 107-17, 2003.

[19] Sokhansanj, B. A., Fitch, J. P., Quong, J. N., and Quong, A. A., "Linear fuzzy gene network models obtained from microarray data by exhaustive search," BMC Bioinformatics, vol. 5, pp. 108, 2004.

[20] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," Mol Biol Cell, vol. 9, No. 12, pp. 3273-97, 1998.

[21] Stuart, D. and Wittenberg, C., "CLB5 and CLB6 are required for premeiotic DNA replication and activation of the meiotic S/M checkpoint," Genes Dev, vol. 12, No. 17, pp. 2698-710, 1998.

[22] Poloumienko, A., Dershowitz, A., De, J., and Newlon, C. S., "Completion of replication map of Saccharomyces cerevisiae chromosome III," Mol Biol Cell, vol. 12, No. 11, pp. 3317-27, 2001.

Disruption of estrogen receptor DNA-binding domain and related intramolecular communication restores tamoxifen sensitivity in resistant breast cancer

Li Hua Wang,[1,7,*] Xiao Yi Yang,[1,7] Xiaohu Zhang,[1] Ping An,[1] Han-Jong Kim,[2] Jiaqiang Huang,[1] Robert Clarke,[3] C. Kent Osborne,[4] John K. Inman,[5] Ettore Appella,[6] and William L. Farrar[2,*]

[1] Basic Research Program, SAIC-Frederick, National Cancer Institute-Frederick, Frederick, Maryland 21702
[2] Cancer Stem Cell Section, Laboratory of Cancer Prevention, National Cancer Institute-Frederick, Frederick, Maryland 21702
[3] Department of Oncology, Lombardi Cancer Center, Georgetown University, Washington, District of Columbia 20007
[4] Breast Center, Baylor College of Medicine, and The Methodist Hospital, One Baylor Plaza, BCM 600, Houston, Texas 77030
[5] Bioorganic Chemistry Section, Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892
[6] Laboratory of Cell Biology, National Cancer Institute, Bethesda, Maryland 20892
[7] These authors contributed equally to this work.
*Correspondence: lhwang@ncifcrf.gov (L.H.W.), farrar@ncifcrf.gov (W.L.F.)

Summary

A serious obstacle to successful treatment of estrogen receptor (ER)-positive human breast cancer is cell resistance to tamoxifen (TAM) therapy. Here we show that the electrophile disulfide benzamide (DIBA), an ER zinc finger inhibitor, blocks ligand-dependent and -independent cell growth of TAM-resistant breast cancer in vitro and in vivo. Such inhibition depends on targeting disruption of the ER DNA-binding domain and its communication with neighboring functional domains, facilitating ERα dissociation from its coactivator AIB1 and concomitant association with its corepressor NCoR bound to chromatin. DIBA does not affect phosphorylation of HER2, MAPK, AKT, and AIB1, suggesting that DIBA-modified ERα may induce a switch from agonistic to antagonistic effects of TAM on resistant breast cancer cells.

Introduction

The selective estrogen receptor modulator (SERM) tamoxifen (TAM), which binds to the estrogen receptor α (ERα) and partially inhibits its activity, is the most prolific therapeutic drug for the treatment of ER-positive breast cancer (Osborne, 1998). Adjuvant therapy studies of TAM show a 40%–50% reduction in the odds of recurrence and reduced mortality. Unfortunately, advanced breast cancers that initially respond well to TAM eventually become refractory to this compound (McDonnell and Norris, 2002; Jordan, 2004; Osborne et al., 2003; Shou et al., 2004).

ER functions in the nucleus as a transcriptional regulator of specific genes (Tsai and O'Malley, 1994). The structural organization of ERα consists of a ligand-independent transcription-activation domain (AF-1 domain), a DNA-binding domain (DBD), a ligand-binding domain (LBD), and a ligand-dependent transactivation domain (AF-2 domain) (Kumar et al., 1987; Ruff et al., 2000). Estrogen binding to ER alters its conformation, triggers receptor dimerization, and directly facilitates binding of the receptor complex to promoter regions of target genes, including sites known as estrogen-responsive elements (ERE), or indirectly through transcription factors such as AP-1 (Kushner et al., 2000). The recruitment of coactivators such as AIB1 and other proteins with acetyltransferase activity helps to unwind the chromatin, allowing transcription to occur (Brzozowski et al., 1997; Glass and Rosenfeld, 2000; Shang et al., 2000; Shiau et al., 1998; Smith et al., 1997). In contrast, the ER conformation induced by the binding of SERMs like TAM favors the recruitment of corepressors NCoR/SMRT and deacetylases that inhibit transcriptional activity in TAM-sensitive breast cancer cells (Keeton and Brown, 2005; Kurebayashi et al., 2000; Mak et al., 1999; Osborne et al., 2003; Shou et al., 2004). However, acquired resistance can be caused by alterations in the ER signal transduction pathway, converting the inhibitory SERM-ERα complex to a growth stimulatory signal (Jordan, 2004).

SIGNIFICANCE

Acquired resistance to antiestrogens is a major challenge to the clinical management of initially endocrine-responsive metastatic breast cancer. We have previously found that electrophilic DIBA and benzisothiazolone derivatives inhibited TAM-sensitive breast cancer cells by preferentially disrupting the vulnerable zinc fingers within the ER DNA-binding domain. Here we describe how DIBA restores the antagonistic action of TAM in resistant breast cancer cells through targeted disruption of the ER DNA-binding domain and its interaction with the proximal N-terminal domain to suppress ligand-dependent and -independent ER transcription and influence the recruitment of cofactor to the ER. These results show that small-molecule modification of the ER zinc finger may alter coactivator/corepressor functions, which are particularly relevant to TAM resistance.

Growing evidence indicates that crosstalk between ER and growth factor receptor signaling pathways (Brockdorff et al., 2003; Ibrahim and Yee, 2005; Osborne et al., 2005), especially the insulin-like growth factor receptor (IGFR) family and the epidermal growth factor receptor (EGFR) family (such as cErbB2 [HER2]), is one of the mechanisms for resistance to endocrine therapy in breast cancer (Schiff et al., 2004). In tumors with abundant ER, AIB1, and HER2, TAM behaves as an ER agonist and stimulates tumor growth (Osborne et al., 2005). High levels of activated AIB1 could reduce the antagonist effects of TAM, especially in tumors that also overexpress the HER2 receptor that activates MAPKs. TAM resistance may also be produced by decreased levels of the corepressor NCoR (Fujita et al., 2003; Lavinsky et al., 1998; Osborne, 1998).

The ER-DBD contains two nonequivalent Cys_4 zinc fingers (Laity et al., 2001; Ruff et al., 2000; Schoenmakers et al., 1999; Wikstrom et al., 1999), which function cooperatively in ER dimerization and DNA binding by stabilizing the secondary and tertiary structure of the ER-DNA complex (Maynard and Covell, 2001; Predki and Sarkar, 1992; Schwabe et al., 1993), leading to ligand-dependent ER transactivation and ER-mediated breast cancer cell and tumor growth. Moreover, interdomain communication between the N-terminal AF-1 domain and DBD of the nuclear receptors helps modulate structure- and ligand-independent functions of receptors (Brodie and McEwan, 2005; Kumar and Thompson, 2003; Shao et al., 1998; Takimoto et al., 2003). We have previously found that electrophilic DIBA and benzisothiazolone derivatives produced anticancer activity in TAM-sensitive human breast cancer cells by preferentially disrupting the vulnerable ER zinc fingers, thus blocking ER DNA binding and transactivation (Wang et al., 2004). Since this anti-breast-cancer strategy targeted ER at the level of its DNA binding, rather than the classical antagonism of estrogen binding, it is relevant to explore whether DIBA has the capacity to inhibit the growth of TAM-resistant breast cancer cells.

In this report, we investigated how DIBA restored the antagonist action of TAM on breast cancer, which was dependent on targeting disruption of the ER DNA-binding domain and its communication with neighboring transcription domains. Moreover, DIBA reduced ER association with coactivator AIB1 and enhanced ER association with corepressor NCoR. These findings provided the proof of principle for a potential for DIBA applicable to TAM-resistant breast cancer.

Results

DIBA suppresses TAM-resistant breast cancer cell growth

First we explored whether DIBA affects estrogen-mediated growth of TAM-resistant breast cancer cells. MCF-7/LCC2 is a selective ER-positive, TAM-resistant cell line (Brunner et al., 1993; Lilling et al., 2000). The specific ER ligand 17β-estradiol (E2) stimulated [^{3}H]thymidine incorporation in MCF-7/LCC2 and its parent MCF-7 cells, but the degree of stimulation in MCF-7/LCC2 is significantly less than that observed in E2-treated MCF-7 cells (Figures 1A and 1B). 4-Hydroxytamoxifen (4-OH-TAM) significantly inhibited MCF-7 cells, with an ED_{50} of 0.1 μM. A low dosage of DIBA enhanced TAM sensitivity, the ED_{50} deceasing 2-fold (0.05 μM) (Figure 1A). The TAM-resistant cell line MCF-7/LCC2 validated with relative resistance; however, a small dosage of DIBA (5 μM) restored 4-OH-TAM sensitivity, achieving over 90% inhibition of E2-driven proliferation at the lowest dosage tested of 4-OH-TAM (0.05 μM). Similarly, DIBA inhibited cell proliferation of MCF-7/HER2-18 (Figure 1C), another TAM-resistant MCF-7 derivative engineered to overexpress HER2 (Benz et al., 1993), and different types of ER-positive and TAM-resistant breast carcinoma cell lines including BT474 (Figure 1D), which expresses ER and is naturally gene amplified for HER2 and AIB1 (Lin et al., 1990; Anzick et al., 1997), and epithelial ZR-75 cells (Figure 1E) (Hoffmann et al., 2004) in a dose-dependent manner. These observations suggested that DIBA effectively restored the antagonist action of TAM on growth of TAM-resistant breast cancer cells.

In TAM-resistant cells, peptide growth factor signaling pathways appear to be important in modifying cell behavior, growth, and survival (Brockdorff et al., 2003; Ibrahim and Yee, 2005). Therefore, we examined whether DIBA impacted TAM-resistant cell growth mediated by stimulation of exogenous peptide growth factors. MCF-7/LCC2 cells (Figure 1F) were stimulated by IGF-1 alone or IGF-1 plus 4-OH-TAM. TAM did not block IGF-1-driven cell proliferation. However, adding DIBA at even 1 μM was sufficient to restore TAM inhibitory functions. These data demonstrated that DIBA also suppressed TAM-resistant cell growth mediated by growth factors.

Efficacy of DIBA on TAM-resistant breast cancer tumor growth in vivo

The in vivo efficacy of the DIBA was tested using nude mice bearing human MCF-7/LCC2 breast carcinoma xenografts. 4-OH-TAM alone did not significantly affect tumor growth. DIBA alone resulted in a dose-dependent inhibition of tumor growth, and a high dose (30 mg/kg) of DIBA reduced tumor volume to almost 50%. Moreover, treatment with 4-OH-TAM plus DIBA diminished tumor to undetectable levels (Figure 2A). Histopathological analysis (Figure 2B) showed a typical hypercellular solid carcinoma invading the dermis and subcutaneum, and the tumor cells had a high nuclear grade with frequent mitosis in the control vehicle (upper panel) or 4-OH-TAM alone-treated mice (middle panel). In contrast, marked reduction in tumor volume, partial encapsulation by fibrous connective tissue, and no significant invasion into surrounding skin tissue were observed in the mice treated with DIBA plus 4-OH-TAM (lower panel). These tumor cells with a low nuclear grade, focal glandular differentiation, and no frequent mitosis or necrosis were seen under higher magnification. No apparent toxicity was observed in liver or kidney in DIBA-treated mice, nor were there any significant changes in body weight gain compared with control mice (data not shown). Therefore, the data demonstrate that DIBA effectively reduces the growth of MCF-7/LCC2 TAM-resistant tumors in mice.

Synergism between DIBA and TAM on cell-cycle progression

Using propidium iodide (PI) staining and fluorescence-activated cell sorting (FACS) analysis, we further evaluated TAM-treated cells within the cell cycle in the presence of DIBA (Figure 2C). E2-treated MCF-7/LCC2 cells showed decreased cells in the G0/G1 phase and an increased percentage of cells in the S and G2/M phases. Cells treated with TAM had a weak inhibitory effect on E2, increasing the percentage of cells in S/G2/M. By

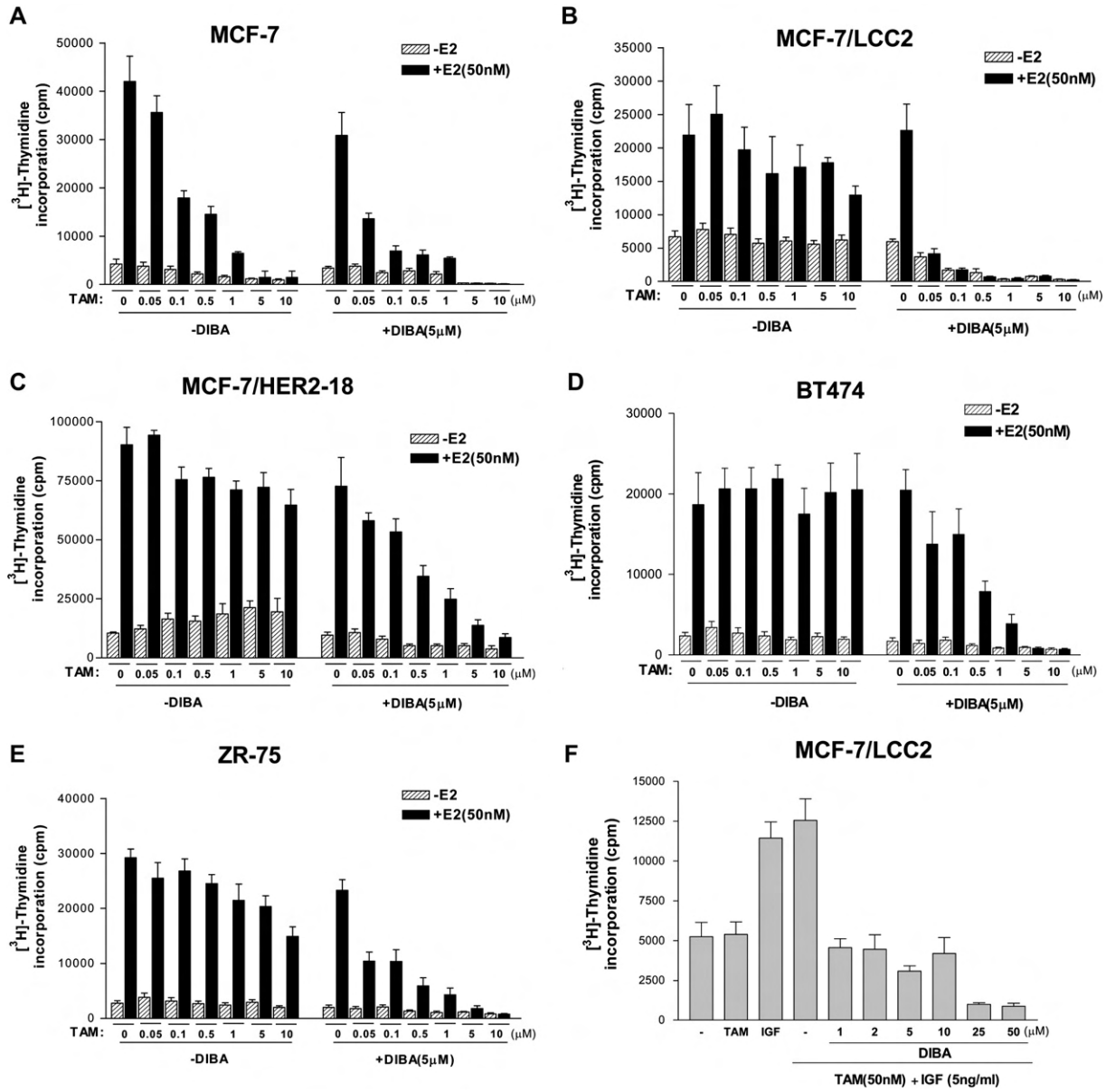


Figure 1. DIBA is a potent inhibitor of TAM-resistant breast cancer cell proliferation

A–E: Proliferation of MCF-7 (**A**), MCF-7/LCC2 (**B**), MCF-7/HER2-18 (**C**), BT474 (**D**), or ZR-75 (**E**) cells was examined by [^{3}H]thymidine incorporation assay. Starved cells were treated with DIBA for 2 hr, stimulated with (filled bars) or without (hatched bars) 50 nM E2, incubated with increasing concentrations of 4-OH-TAM, and analyzed 48 hr later. Data shown represent mean ± SEM.

F: Proliferation of starved MCF-7/LCC2 cells induced by 50 nM 4-OH-TAM or 5 ng/ml IGF-1 was also examined after treatment with increasing concentrations of DIBA. Data shown represent mean ± SEM.

contrast, in cells cotreated with E2 plus ICI 182780, the changes in cell-cycle status and growth induced by E2 were significantly inhibited. In the presence of DIBA combined with TAM, cell-cycle phase distribution induced by E2 shows a significant increase (from 60.7% to 81.7%) of cells in the G0/G1 phase, a decrease (from 30.1% to 6.4%) in the S phase, a decrease (from 7.9% to 5.4%) in the G2/M phase, and an increase (from 1.2% to 6.5%) in the sub-G1 phase. Also, DIBA enhanced the inhibitory effect of ICI 182780 on E2-stimulated cell growth. The FACS data further confirmed that DIBA restored the antagonist action of TAM on cell proliferation of TAM-resistant breast cancer cells analyzed by the [^{3}H]thymidine incorporation assay.

ER is necessary for synergism between DIBA and TAM

To determine whether targeted disruption of ER is necessary for DIBA to suppress cell growth of TAM-resistant breast cancer cells, we used BT474, an ER-positive but TAM-resistant breast cancer cell line, as a model system to examine the effect of depletion of ER on DIBA inhibition of cell growth of TAM-resistant cells (Figure 3A). The ER expressed in these cells was knocked down by using ERα-siRNA. The decreased level of ER was confirmed by western blot (Figure 3A, inset). Under identical conditions, DIBA rendered TAM inhibition on parent ER-positive cells, but was not able to sensitize TAM's suppression of growth of ER-depleted breast cancer cells. These data suggest that

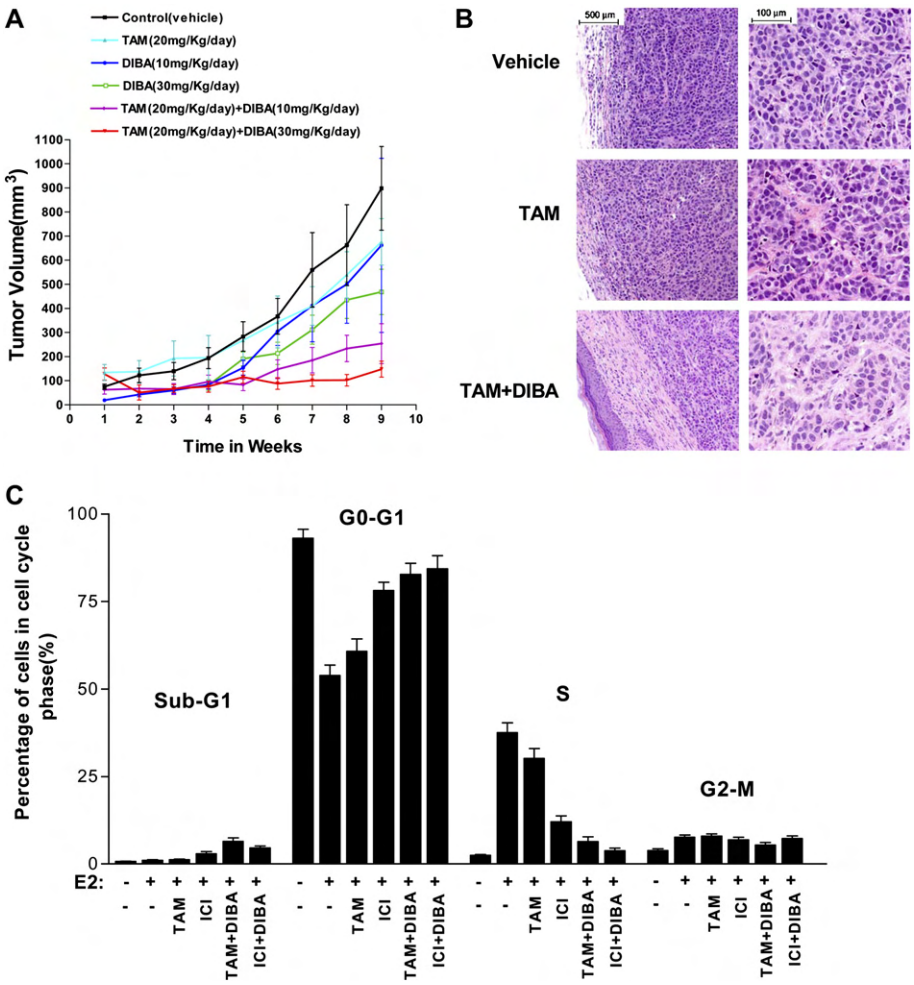


Figure 2. Synergism inhibition between DIBA and TAM on in vivo tumor growth and cell-cycle progression

A: Dose-dependent effect of DIBA and 4-OH-TAM on growth of MCF-7/LCC2 tumor in mice. Data shown represent mean ± SEM (n = 10 mice per group).

B: Morphology of MCF-7/LCC2 tumors treated with vehicle (upper panel), 4-OH-TAM at 20 mg/kg/day (middle panel), or 4-OH-TAM at 20 mg/kg/day plus DIBA at 30 mg/kg (lower panel).

C: MCF-7/LCC2 cells were synchronized by serum starvation, pretreated with 5 μM DIBA, and then stimulated with 50 nM E2, 50 nM TAM, or 1 μM ICI 182780. Cell-cycle distribution was examined by PI staining and FACS analysis. The results represent three independent experiments (mean ± SEM).

inhibition of DIBA on growth of TAM-resistant breast cancer cells depends on ER.

DIBA inhibits ER binding to DNA

To clarify whether the DIBA alters estrogen- or TAM-bound ER's ability to bind to its cognate ERE in TAM-resistant breast cancer cells, we performed electrophoretic mobility shift assays (EMSA) using nuclear extracts obtained from MCF-7/LCC2 cells (Figure 3B). E2- or 4-OH-TAM-treated cells displayed considerable ERE DNA-binding complexes, which could be partially supershifted with anti-ER, but not normal rabbit serum, confirming the specificity of these binding complexes. DIBA significantly decreased (80%) the E2- or TAM-induced ERE DNA-binding activity. In a similar experiment on androgen receptor (AR) in MCF-7/LCC2 (Figure 3C), DIBA did not inhibit AR DNA-binding activity.

Next, we examined whether DIBA affects ER binding to probes containing the AP-1, a nontypical repeat element. Estrogen or TAM induced substantial AP-1 binding activity (Figure 3D). The complexes were mostly supershifted with anti-Jun or anti-Fos antibodies. Anti-ER antibody just marginally decreased such complexes, suggesting that a low amount of ER may be bound to AP-1 sites under these conditions/cells. Moreover, DIBA did not display an inhibitory effect on E2- or TAM-stimulated AP-1-binding activity, possibly because ER binding to DNA is not required for its activity through the nonclassical AP-1 pathway (Jakacka et al., 2001; Webb et al., 1999). These data further support the specificity of DIBA influencing ER binding to DNA.

DIBA blocks occupancy of estrogen target gene promoters by ERα

We further used chromatin immunoprecipitations (ChIP) to directly assess whether DIBA impacts ERα binding to promoters of estrogen target genes. The presence of the specific promoters in the chromatin immunoprecipitates was analyzed by semiquantitative PCR by using specific pairs of primers spanning the estrogen-responsive regions in the pS2, c-Myc, and cathepsin D (CATD) gene promoters (Figure 3E). Stimulation with E2 and TAM dramatically increased ERα's occupancy of the above three promoters. DIBA remarkably decreased such occupancy of ERα to the target gene DNA sequences in chromatin. By contrast, ERα did not show any interaction with the distal promoter region (−3351 to −3551) of pS2 promoter. These results suggested DIBA directly influences the ability of ERα to bind to ERE in the promoter of target genes.

We also used ChIPs to examine whether DIBA affects ER binding to AP-1 site in a nontypical manner (Figure 3F). Stimulation with E2 and TAM induced a dramatic increase in the occupancy by c-*fos* or ERα of the AP-1 site, but not in the

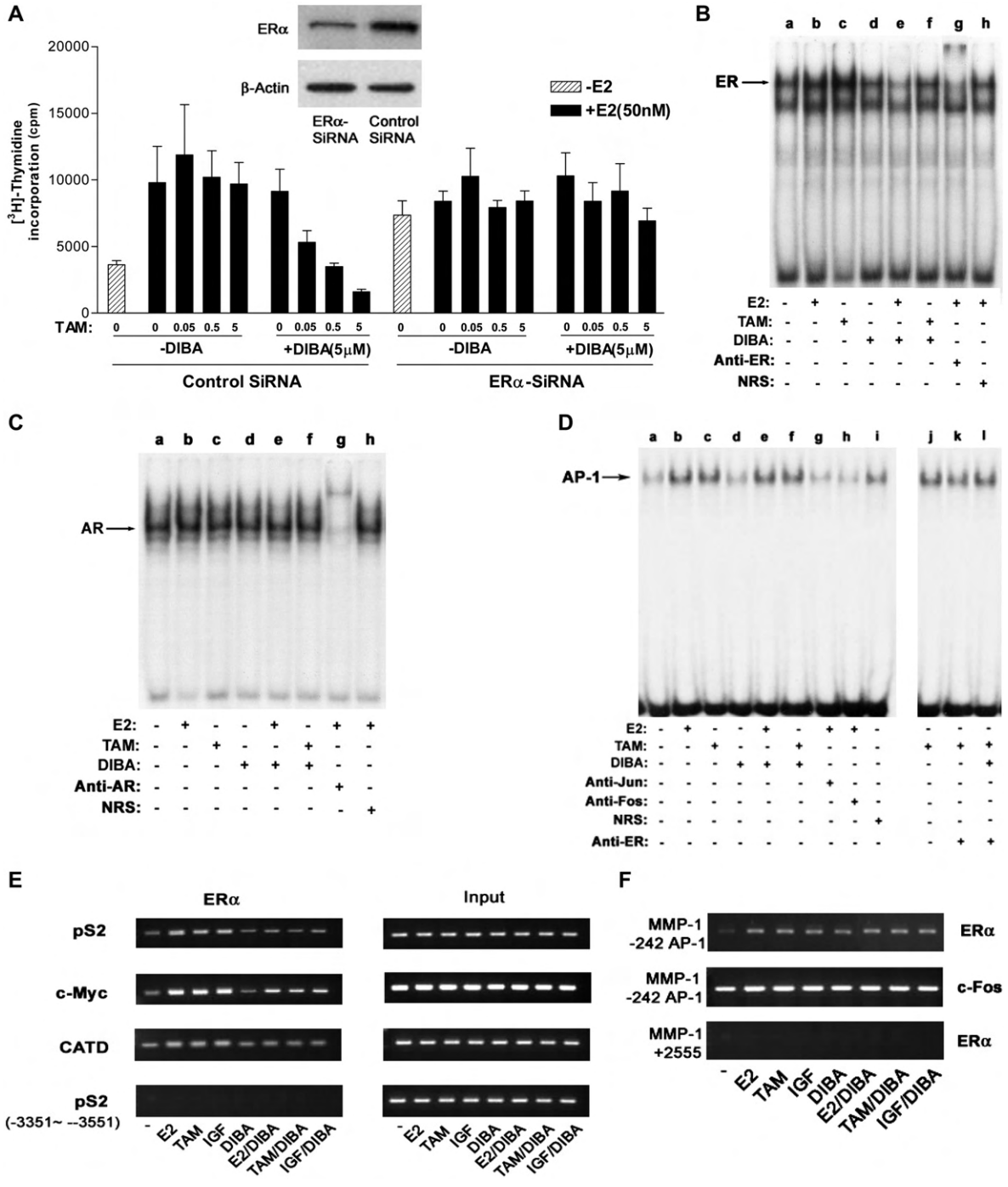


Figure 3. ER is necessary for DIBA to sensitize TAM inhibition

A: siRNA-mediated knockdown of ERα alters DIBA-mediated TAM inhibition of resistant cell growth. BT474 cells were transduced with ERα-siRNA or control vector and incubated for 96 hr. Levels of ERα expression were examined by western blotting (inset). Proliferation of the above transfected cells treated with DIBA and 4-OH-TAM in the presence of E2 was assayed by [^{3}H]thymidine incorporation. Data shown represent mean ± SEM.

B–D: DIBA inhibits E2-induced ERE (**B**), but not ARE (**C**) or AP-1 (**D**), DNA binding. MCF-7/LCC2 cells were treated with or without 5 μM DIBA for 2 hr, then stimulated with medium (−), 50 nM E2, or 50 nM 4-OH-TAM (+) for 20 min. Nuclear extracts were incubated in the absence of antibody, αER, αAR, αJun, αFos, or normal rabbit serum (NRS) in combination with ^{32}P-labeled oligonucleotide probes. Arrows indicate migrational location of each nonsupershifted ER, AR or AP-1 DNA complex.

E: The recruitment of ERα to the promoters of estrogen-responsive genes. MCF-7/LCC2 cells were treated with or without 5 μM DIBA for 2 hr, then stimulated with E2, 4-OH-TAM, or IGF-1 for 40 min. Soluble chromatin was prepared and immunoprecipitated with anti-ERα. The final DNA extractions were amplified using pairs of primers that cover the regions of pS2, CATD, and c-Myc gene promoters, as indicated. The distal region (approximately −3351 to −3551) of the pS2 gene promoter was examined for the presence of ERα (bottom row).

F: The recruitment of ERα to the promoter of an estrogen-induced AP-1-dependent gene MMP-1. Soluble chromatin was immunoprecipitated with antibodies against ERα or c-Fos. The final DNA extractions were amplified using pairs of primers that cover the AP-1 site as indicated or the non-AP-1-specific site (approximately +2555) of the MMP-1 gene promoter.

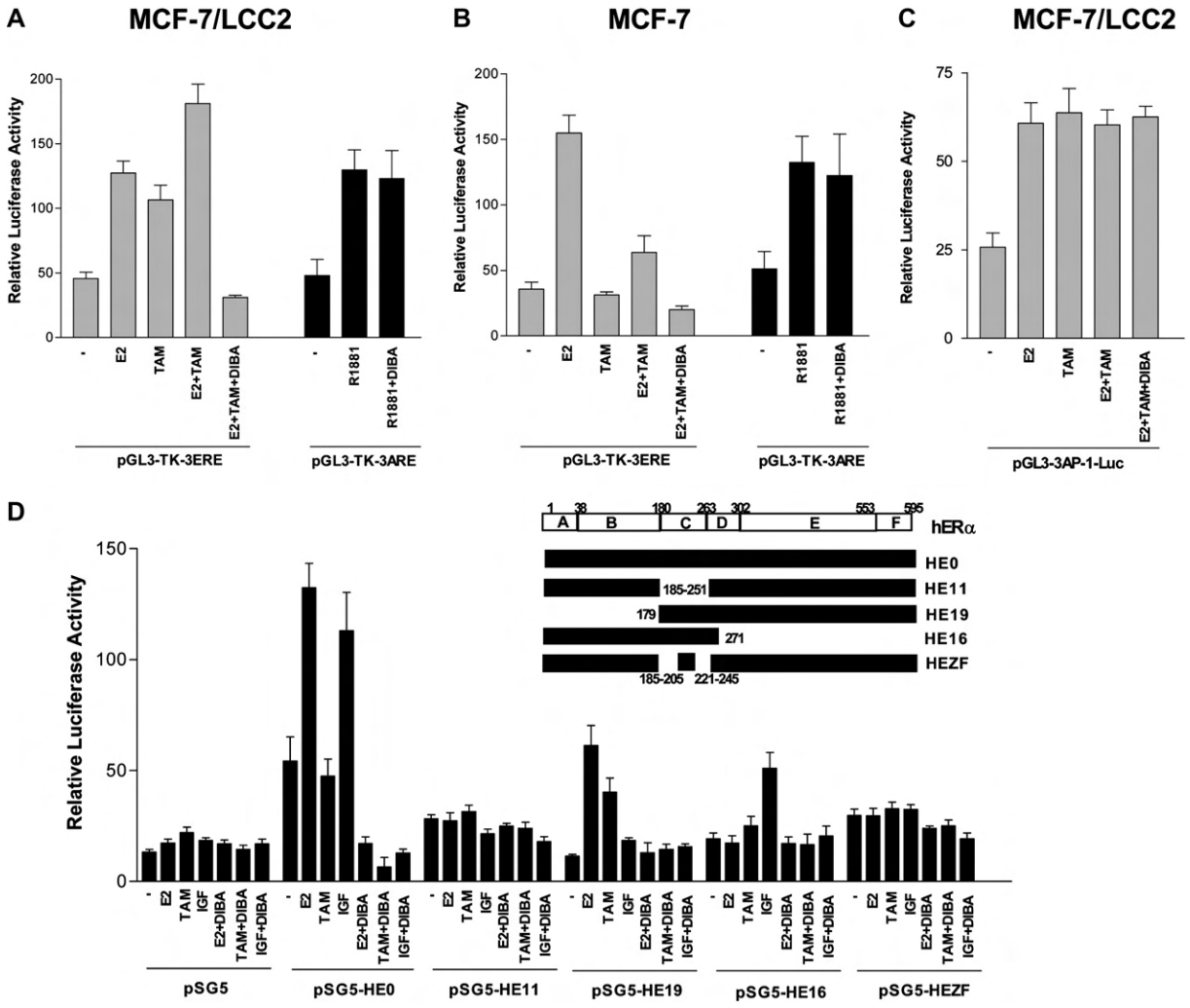


Figure 4. DIBA inhibits ERE transactivation

A–C: MCF-7/LCC2 or MCF-7 cells were transfected with a pGL3-TK-ERE luciferase, pGL3-TK-ARE luciferase, or pGL3-AP-1 luciferase construct, respectively. After addition of 4-OH-TAM (50 nM) and/or DIBA (5 μM) for 2 hr, cells were stimulated with or without 50 nM E2 or 100 nM R1881 for 16 hr. Luciferase activity of lysed cells was measured and normalized. Data shown represent mean ± SEM.

D: MDA-MB-468 cells were transfected with a wild-type ER (pSG5-HE0), a series of human ER deletion mutants including pSG5-HE11, pSG5-HE19, pSG5-HE16, pSG5-HEZF, or pSG5 control plasmids and a pGL3-TK-ERE luciferase reporter. After 24 hr, the transfected cells were treated with DIBA, E2, 4-OH-TAM, and IGF-1 for an additional 24 hr. Luciferase activity of lysed cells was measured and normalized. Data shown represent mean ± SEM.

non-AP-1-specific site in promoter of matrix metalloproteinase 1 (MMP-1), an estrogen-induced/AP-1-dependent gene promoter containing AP-1 sites but no ERE sequences (DeNardo et al., 2005). DIBA did not affect such occupancy of c-*fos* or ERα, consistent with the observation by EMSA.

DIBA inactivates ligand-dependent ERE transactivation

To determine whether DIBA might affect TAM-mediated ER transcription in TAM-resistant breast cancer cells, we tested transactivation of MCF-7/LCC2 (Figure 4A) and MCF-7 (Figure 4B) cells transfected with the ERE-luciferase reporter gene. E2 activated ERE transactivation in both cell lines. TAM alone suppressed E2-induced ERE transactivation in MCF-7 cells, whereas it increased ERE transactivation in MCF-7/LCC2 cells. DIBA significantly reduced ERE transactivation stimulated by 4-OH-TAM and E2 in MCF-7/LCC2 cells. By contrast, DIBA did not affect androgen-responsive element (ARE) transactivation mediated by R1881 in both MCF-7 and MCF-7/LCC2 cells (Figures 4A and 4B). Furthermore, DIBA did not inhibit transactivation of AP-1-luc (Figure 4C). These data indicate that DIBA selectively suppresses TAM-stimulated ER DNA binding and subsequent ERE transactivation.

To further validate the target specificity of DIBA on ligand-dependent ERE transcription in TAM-resistant breast cancer cells, we cotransfected the wild-type human ERα (HE0), a series of human ER deletion mutants (Kumar et al., 1987) including HEZF (ER depleted of zinc finger domains, ER-ΔZF), HE11 (ER depleted of DBD, ER-ΔDBD), HE19 (ER depleted of A/B regions but containing DBD and AF-2 domain, ER-ΔA/B), HE16 (ER depleted of D/E/F regions, ER-ΔD/E/F), or pSG5 control expression plasmid and the ERE-luciferase reporter gene into the ER-negative MDA-MB-468 cells. As shown in Figure 4D, overexpression of ERα, compared to pSG5 control, remarkably resulted in ERE transactivation. E2 strongly activated ERE

transactivation, whereas TAM did not show significant inhibition on such transactivation. Also, induction of ERE transactivation by E2 was observed in the cells overexpressed by ERα mutant HE19 containing completed AF-2 domain and DBD. Deletion of zinc finger domains (HEZF) or the entire DBD (HE11) resulted in decreasing ERE transactivation stimulated by E2, suggesting that zinc fingers in DBD are required for ligand-dependent ERE transactivation. DIBA significantly enhanced TAM inhibition in the wild-type ERα- or a mutant HE19 (ER-ΔA/B)-overexpressing cells, but not in HE11 (ER-ΔDBD)- or HEZF (ER-ΔZF)-overexpressing or control cells. Since DIBA, as a zinc finger inhibitor, has been demonstrated to preferentially disrupt the ER DNA-binding domain, the inhibitory effect of DIBA on ligand-induced ERE transcription in TAM-resistant breast cancer cells may be related to interruption of zinc finger domains within ER-DBD.

DIBA reduces ligand-independent ERE transactivation

The ligand-independent ERE transcription was also measured in the above wild-type ER or mutant-transfected MDA-MB-468 cells. As shown in Figure 4D, in the case of the wild-type ER (HE0)-overexpressing cells, IGF-1 strongly induced ERE transactivation. Deleting zinc finger domains or the entire DBD decreased ERE transactivation stimulated by either E2 or IGF-1, even though this mutant contains completed AF-1 and AF-2 domains. However, induction of ERE transactivation by IGF-1 was observed in the cells overexpressed by the ER mutant HE16 containing a completed N-terminal A/B domain and DBD, suggesting that DBD is required for both ligand-dependent and -independent ERE transactivation. IGF-1's activation of ERE transcription is not only dependent on the AF-1 domain itself, but is also mediated through the interaction between DBD and AF-1 domains, consistent with previous observations that long-range allosteric communication occurs in two separated domains of the androgen receptor (Brodie and McEwan, 2005), glucocorticoid receptor (Kumar et al., 1999), and progesterone receptor (Bain et al., 2000).

Moreover, DIBA blocked ERE transactivation stimulated by IGF-1 in the cells overexpressed by a wild-type ER or the ER mutant (HE16) containing A/B/C domains. Such an inhibitory effect of DIBA was not observed in the ER mutants HE11 (ER-ΔDBD)- and HEZF (ER-ΔZF)-transfected cells, indicating that inhibition of DIBA on the "steroid-independent activation" of ER by growth factor signals was related to DBD-mediated intramolecular communication with the AF-1 domain, which may also be involved in DIBA functionally suppressing TAM-resistant breast cancer cells.

DIBA decreases the TAM-bound ER association with AIB1

Activated AIB1 probably translocates to nucleus (Schiff et al., 2004), where it can interact with ER; therefore, we utilized a coimmunoprecipitation experiment to analyze whether DIBA impacts the ERα interaction with AIB1. Cell extracts were immunoprecipitated with an anti-ERα-specific antibody; immunoprecipitates were developed on western blots with anti-AIB1 (upper panel) or anti-ERα (lower panel). In MCF-7/LCC2 cells (Figure 5A), the AIB1 can be coprecipitated with ERα in cells treated with E2, 4-OH-TAM, or IGF-1, indicating that a direct protein-protein interaction occurs between nuclear receptor ERα and its coactivator AIB1 upon addition of E2, 4-OH-TAM, and IGF-1. Notably, DIBA significantly decreased such ER interaction with AIB1. In contrast, E2, but not TAM, induced this association between ERα and AIB1 in MCF-7 cells. These data support that the effect of DIBA on TAM-resistant MCF-7/LCC2 cells may be through dissociation of the coactivator AIB1 complexes from TAM-bound ERα.

DIBA increases association of TAM-bound ERα with NCoR

Several lines of evidence indicate that the nuclear receptor corepressor (NCoR) complex mediates the inhibitory effects of TAM (Keeton and Brown, 2005). Thus, we examined whether DIBA affects NCoR modulation of the response of ERα to TAM by using a coimmunoprecipitation experiment. Cell extracts were immunoprecipitated with an anti-NCoR specific antibody; immunoprecipitates were developed on western blots with anti-ERα (Figure 5B). In control MCF-7 cells, TAM induced the association between ERα and NCoR while E2 did not affect it, which may mediate the antagonistic effect of TAM on its sensitive cells. In MCF-7/LCC2 cells, a little ERα can be coprecipitated with NCoR, suggesting that a weak constitutive interaction occurs between nuclear receptor ERα and NCoR, which is consistent with the previous observations that interactions of ERα with NCoR in vitro appear to occur regardless of the ligand state of the receptor (Smith et al., 1997; Voss et al., 2005). Although E2 and IGF-1 significantly decreased such interaction, TAM alone did not increase it. DIBA remarkably increased NCoR association with ERα in the presence of E2, IGF-1, and TAM, suggesting that effect of DIBA on TAM-resistant MCF-7/LCC2 cells may also occur through association of the corepressor NCoR complexes with ERα.

DIBA mediates chromatin-associated recruitment of ERα and cofactors

To examine whether interaction between ERα and AIB1 or ERα and NCoR is chromatin associated, we performed ChIP assays of ER followed by the Re-ChIP analysis of either AIB1 or NCoR, analyzing the assembly of ERα-cofactor complex components on a well-characterized estrogen-responsive pS2 promoter (Figure 5C). The soluble chromatin derived from MCF-7/LCC2 cells was subjected to ChIP with ERα antibodies; subsequently, the released immune complexes were divided into two aliquots for the Re-ChIP using AIB1 antibodies or NCoR antibodies. The same Re-ChIP was also performed on the unbound supernatant fractions from the primary immunoprecipitation. The ChIP assay of ERα antibodies showed that strong binding of ERα to the pS2 promoter was induced by E2 or TAM. DIBA significantly decreased E2- or TAM-occupied ERα binding to the estrogen-responsive DNA sequences in the pS2 promoter. The Re-ChIP assay using AIB1 antibodies illustrated that E2 or TAM induced occupancy of the pS2 promoter by ER and the coactivator AIB1. However, the Re-ChIP assay using NCoR antibodies showed that a marginal recruitment of the NCoR occurred in the absence of ligand, while stimulating E2 or TAM abolished such promoter occupancy by ERα-NCoR complexes, indicating that interactions between ERα-AIB1 and between ERα-NCoR are chromatin associated. After DIBA treatment, there were very low levels of E2- or TAM-induced recruitment of ERα-AIB1 and ERα-NcoR complexes to chromatin. Combined with the data obtained from coimmunoprecipitation experiments (Figures 5A and 5B), these results suggested that DIBA-induced changes in ERα association with cofactors led to inhibition of ERα binding to DNA, in

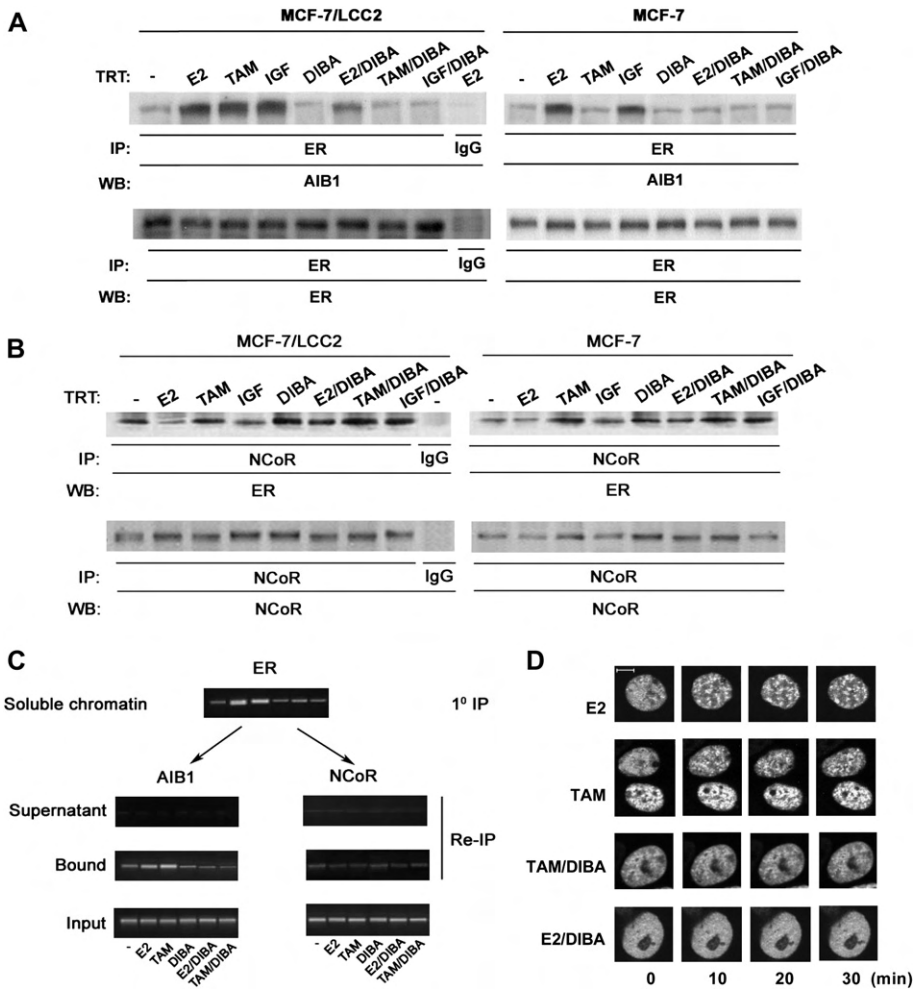


Figure 5. Effect of DIBA on ERα association with AIB1 or NCoR

A and B: ERα association with cofactors assayed by coimmunoprecipitation. MCF-7/LCC2 or MCF-7 cells were treated with or without DIBA for 2 hr, and then stimulated with E2, 4-OH-TAM, or IGF-1 for 24 hr before lyses. **A:** Western blotting analysis with anti-AIB1 (upper panel) or anti-ERα (lower panel) was performed on anti-ERα immunoprecipitates. **B:** Western blotting analysis was performed with anti-ERα (upper panel) or anti-NCoR (lower panel) on anti-NCoR immunoprecipitates.

C: Recruitment of ERα and cofactors assayed by ChIP-Re-ChIP. Soluble chromatin was immunoprecipitated with antibodies against ERα (1° IP). The supernatant was collected and reimmunoprecipitated with antibodies against AIB1 or NCoR (Supernatant Re-IP). Similar reciprocal Re-IPs were also performed on complexes eluted from the 1° IPs (Bound Re-IP).

D: Time course of GFP-ERα redistribution. MCF-7/LCC2 cells were transiently transfected with pEGFP-C2-hERα. Live cells expressing GFP-ERα were pretreated with vehicle or DIBA for 2 hr, followed by stimulation with 50 nM E2 or 50 nM TAM. Time courses of GFP-ER distribution were analyzed at 10 min intervals. Scale bar, 5 μm.

turn blocking transcription of target genes, which aided the synergism between DIBA and TAM.

Since cofactor association can influence ERα cellular localization, we used a transcriptionally active green fluorescent protein-ERα chimera (GFP-ERα) to examine whether DIBA affects ERα cellular distribution. MCF-7/LCC2 cells were transiently transfected with pEGFP-C2-hERα, and live cells expressing GFP-ERα were analyzed at 10 min intervals under confocal laser scanning microscopy. Without ligand, GFP-ERα was observed only in the nucleus, excluding the nucleolus, with a diffuse distribution. Upon adding E2, GFP-ERα was dramatically redistributed from a reticular to punctate pattern within the nucleus (Figure 5D). A similar reorganization occurred with TAM. In the cells pretreated with DIBA, neither E2 nor TAM produced the above apparent subnuclear redistribution patterns. These results demonstrated that DIBA inhibited E2- or TAM-induced ERα nuclear distribution.

DIBA dephosphorylates ERα at serine-167, but not serine-118

The human ERα AF-1 function is potentiated by the phosphorylation of serine residues of human ERα A/B domain after stimulation with its ligands and nonsteroidal growth factors (EGF and IGF-1) (Lannigan, 2003; Yamashita et al., 2005). We thus investigated whether DIBA may modulate ERα phosphorylation by using site-specific antiphosphoserine antibodies against ERα at Ser-118 or Ser-167 (Figure 6A). E2, 4-OH-TAM, and IGF-1 stimulated Ser-167 phosphorylation, whereas there was no significant difference in the level of phosphorylation of ERα at Ser-118 in MCF-7/LCC2 cells with the above treatments. While DIBA inhibited phosphorylation of ERα at Ser-167 induced by all stimuli, it affected neither Ser-118 phosphorylation nor the expression of ERα. It has been demonstrated that ERα phosphorylation at Ser-167, but not at Ser-118, conferred DNA binding and transcriptional activation (Joel et al., 1998) as well as TAM resistance (Campbell et al., 2001). Since the structure of the N-terminal AF-1 domain appears to be influenced by the DBD (Graham et al., 2000), and DIBA selectively reacts with zinc finger of ER-DBD, it seems likely that DIBA may interfere with phosphorylation of Ser-167 in AF-1 proximal to the DBD site through intramolecular communication, concurring with the above observation on the effect of DIBA in ligand-independent ERE transactivation (Figure 4D), and may also contribute to DIBA sensitizing the resistant cells to TAM.

DIBA does not affect expression and phosphorylation of AIB1 and MAPK

Since ER coactivator AIB1, like ER itself, is phosphorylated and activated by different signaling kinases, including the p42/44 MAPK, which can be activated by HER2 (Font de Mora and

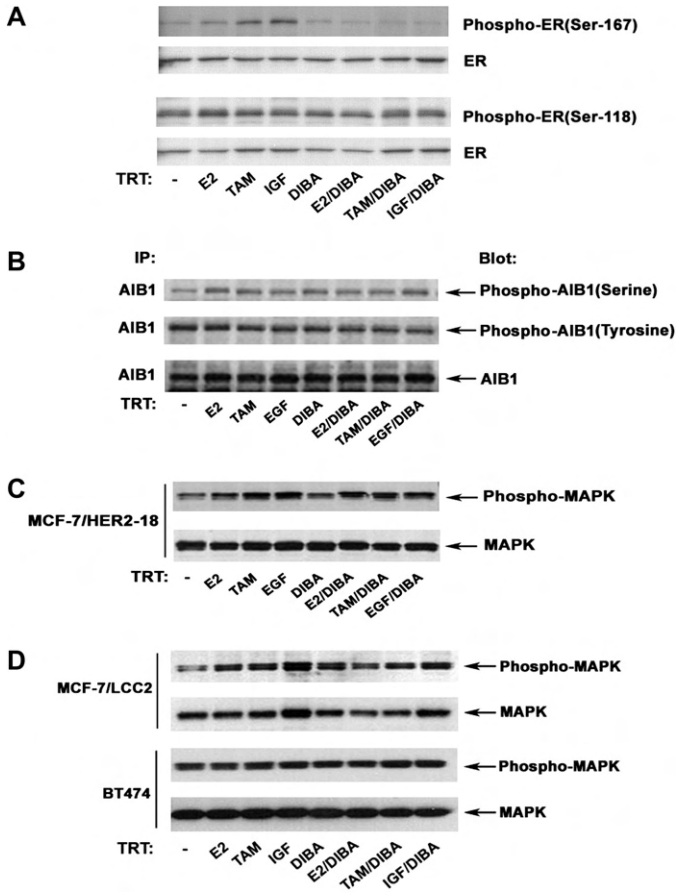


Figure 6. Effect of DIBA on expression and phosphorylation of ERα, AIB1, and MAPK

A: MCF-7/LCC2 cells were treated with or without DIBA for 2 hr, then simulated with E2, 4-OH-TAM, or IGF-1 for 20 min before lyses. Western blotting analysis was performed with anti-phospho-ER (Ser-167 or Ser-118) or anti-ER.
B: MCF-7/HER2-18 cell lysates were immunoprecipitated with anti-AIB1. Immunoprecipitates were blotted with anti-phosphoserine (upper panel), anti-phosphotyrosine (middle panel), or anti-AIB1 (lower panel).
C and D: Cell lysates of MCF-7/HER2-18 (**C**), MCF-7/LCC2, and BT474 (**D**) were analyzed with anti-phospho-MAPK for blot (upper panel) or anti-MAPK (lower panel) for re-blot.

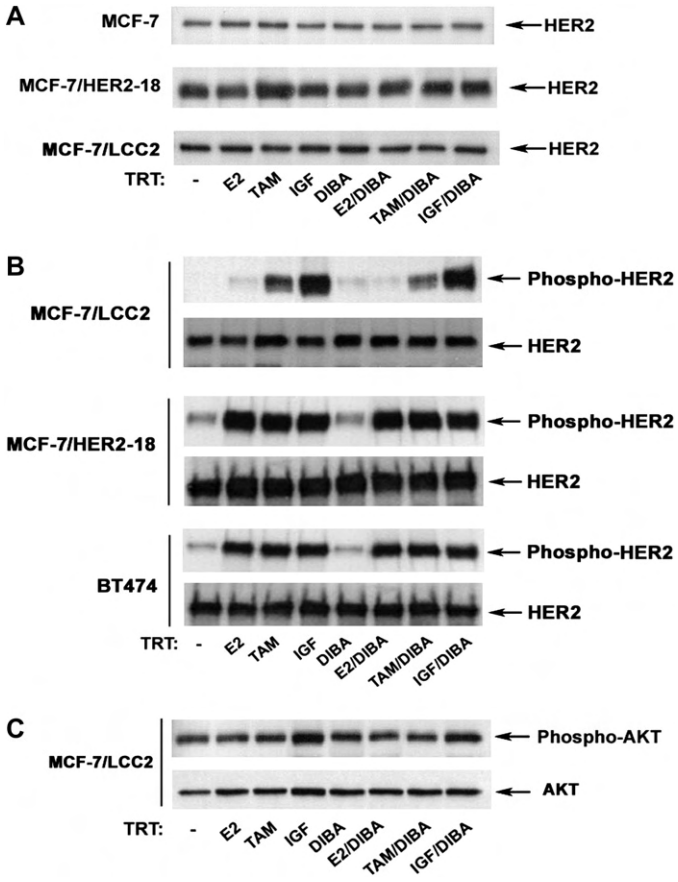


Figure 7. Effect of DIBA on expression and phosphorylation of HER2 and AKT

A: MCF-7, MCF-7/LCC2, and MCF-7/HER2-18 cells were treated with DIBA for 2 hr and then stimulated with E2, 4-OH-TAM, or IGF-1 for 20 min. HER2 expression was analyzed by western blotting with anti-HER2.
B: MCF-7/LCC2, MCF-7/HER2-18, and BT474 cells were treated as described in **A**, except that the antibodies were anti-phospho-HER2 for blot (upper panel) or anti-HER2 for re-blot (lower panel).
C: MCF-7/LCC2 cells were treated as described in **A**, except that the antibodies were anti-phospho-AKT for blot or anti-AKT for re-blot.

Brown, 2000), we examined whether DIBA attenuates phosphorylation of AIB1 in TAM-resistant cells (Figure 6B). MCF-7/HER2-18 cell extracts were immunoprecipitated with an anti-AIB1 specific antibody; the immunoprecipitates were developed on western blots with anti-phosphoserine (upper panel), anti-phosphotyrosine (middle panel), or anti-AIB1 (lower panel). The phosphorylation of serine, but not tyrosine, of AIB1 could be observed in cells stimulated with E2, 4-OH-TAM, and EGF. However, DIBA did not affect such phosphorylation, indicating that DIBA inactivates ER Ser-167 phosphorylation, but does not affect expression and phosphorylation of AIB1, possibly due to AIB1's lacking a zinc finger, although the signaling from the EGFR/HER2 family activates ER and AIB1 by the p42/44 MAPK.

We further examined whether DIBA disrupted phosphorylation of MAPK in different TAM-resistant breast cancer cell lines. Figure 6C shows the same pattern for the phosphorylation of MAPK as that for AIB1 in MCF-7/HER2-18 cells. Similar results were observed in MCF-7/LCC2 and BT474 cells (Figure 6D); the ratio of phospho-MAPK to total MAPK was not significantly changed after the treatment with DIBA.

DIBA does not influence expression and phosphorylation of HER2 and AKT

Since overexpression of HER2 and high levels of phosphorylated AKT may also contribute to TAM resistance (Gutierrez et al., 2005; Osborne et al., 2003), we examined whether DIBA disrupted expression and phosphorylation of HER2 and AKT in TAM-resistant breast cancer cell lines. Compared to MCF-7 cells, MCF7/HER2-18, MCF-7/LCC2, and BT474 cells expressed a considerable level of HER2 (Figures 7A and 7B). There are no significant changes in HER2 expression in DIBA-treated cells. Moreover, even though TAM and IGF-1 induced remarkable phosphorylation of HER2 (Stoica et al., 2000), DIBA did not affect it (Figure 7B), nor did DIBA significantly affect TAM- or IGF-1-stimulated phosphorylation of AKT (Figure 7C) in MCF-7/LCC2 cells. These results indicate the inhibitory effect of DIBA on TAM-resistant cell proliferation is not based on inactivation of HER2, MAPK, and PI3-K/AKT.

Discussion

ER plays a major role in many cases of breast cancer and apparently contributes to the growth of TAM-responsive, acquired TAM-resistant, and de novo ER-positive resistant models (Gee et al., 2005). By using siRNA to deplete ER of BT474, an ER-positive but TAM-resistant breast cancer cell line, we found that DIBA rendered TAM inhibition on parent ER-positive cells, but not on ER-depleted breast cancer cells (Figure 3A), suggesting that DIBA function on TAM-resistant breast cancer cells is ER dependent. We previously discovered that DIBA preferentially disrupted the vulnerable zinc fingers of the ERα DNA-binding domain, thus blocking ER DNA binding and transactivation (Wang et al., 2004). In this study, we used chromatin immunoprecipitations to directly assess ER binding to DNA on estrogen target genes. Stimulation with E2 and TAM dramatically increased ER's occupancy of the pS2, c-Myc, and CATD gene promoters. DIBA remarkably decreased such occupancy of ER to the target DNA sequences in chromatin (Figure 3E). These results suggested DIBA directly influences the ability of ER to bind to DNA, consistent with the data obtained from EMSA (Figure 3B). Moreover, DIBA resulted in inhibition of ligand-dependent and -independent ERE transactivation (Figure 4D). Therefore, targeted disruption of ER is necessary for DIBA, a zinc finger inhibitor, to sensitize TAM inhibition of resistant breast cancer cells through interfering with ER DNA binding and subsequent ERE transactivation.

Nuclear receptor function is modulated by transcriptional coregulators (Klinge et al., 2004; McKenna and O'Malley, 2002; Shang et al., 2000; Shang and Brown, 2002; Tikkanen et al., 2000). The relative level of these coactivators and corepressors might determine the balance of agonist and antagonist properties of TAM. Here, we used coimmunoprecipitation to clarify that DIBA decreased physical association of TAM-bound ER with its coactivator AIB1 (Figure 5A), whereas it increased ER interaction with its corepressor NCoR (Figure 5B). Moreover, ChIP experiments of ERα followed by either Re-ChIP of AIB1 or NCoR also showed that effect of DIBA on E2- or TAM-induced association between ER and AIB1 and dissociation between ERα and NCoR in the level of chromatin (Figure 5C), suggesting that DIBA-mediated changes in ERα interaction with cofactors resulted in blockage of TAM-bound ERα binding to targeted gene promoter and transcription. Notably, the ER-cofactor association caused by DIBA further influences TAM-bound ER nuclear distribution (Figure 5D), indicating other functional changes of ERα may have with chromatin on/off rates or shuttling. The above molecular mechanisms, by which the synergism between DIBA and TAM impacted ER activity, contributed to DIBA restoring TAM's antagonist action on TAM-resistant breast cancer cells (Figures 1 and 2). It may be important to note that in our previous report, the ED_{50} for DIBA ejection of zinc from recombinant ERα was 25 μM (Wang et al., 2004). Here we show effects on disrupting ERα/AIB1 or enhancing ERα/NCoR functions at 5-fold less, 5 μM, suggesting a range of molecular effects on ERα functions.

Several peptide growth factors and their intracellular signaling kinases, notably MAPK and AKT (Albanell and Baselga, 2001), have been shown to mediate cell proliferative responses and phosphorylate ERα on various AF-1 residues, promoting ERα transcriptional activity in a ligand-independent manner (Martin et al., 2000). In the case of TAM-resistant cells, we observed that exogenous IGF-1 stimulated phosphorylation of MAPK and AKT as well as ERα. Although DIBA did not affect phosphorylation of ERα at Ser-118, MAPK, or AKT, DIBA markedly inhibited phosphorylation of ERα at Ser-167 (Figure 6A), suggesting that inhibitory effects of DIBA on a powerful functional crosstalk engaged by the IGF-1 and ER pathways may occur through dephosphorylating ERα at Ser-167. Thus, DIBA disruption of ER zinc fingers resulted in not only perturbing DBD-dependent ERE transactivation (Figures 4A–4D), but also interfering with intramolecular communication between DBD and the N-terminal AF-1 domain (Figure 4D) to downregulate phosphorylation of Ser-167 (Figure 6A) induced by nonestrogenic stimulation in TAM-resistant breast carcinoma cells.

Overexpression of HER2 and high levels of phosphorylated AKT or ERK1/2 MAPK may also contribute to TAM resistance. MCF-7 cells stably transfected with HER2 (MCF-7/HER2-18) are de novo resistant to TAM, in contrast to their low-expressing, responsive MCF-7 counterparts (Benz et al., 1993; Konecny et al., 2003; Shou et al., 2004). Importantly, EGFR/HER2 signaling remains dependent on the ER in MCF-7/HER2-18 cells, as evidenced by their retained sensitivity to estrogen deprivation (Shou et al., 2004). DIBA did not inhibit phosphorylation of HER2, MAPK, and AKT (Figures 6 and 7) or of the coactivator AIB1 (Figure 6B). Moreover, DIBA did not display suppression of estrogen- or TAM-induced DNA binding (Figures 3D and 3F) and transactivation for AP-1 (Figures 3F and 4C), a nontypical ER-binding site. These results suggested that nongenomic actions of ERα may be not involved in the synergism between DIBA and TAM. However, this idea can not be totally excluded; most of the data suggest that DIBA blocks classical genomic sites of ERα.

In conclusion, DIBA resulted not only in inhibition of ligand-dependent ERα DNA binding and transcription, but also in effects on ligand-independent ERE transactivation. Of particular importance was the synergism between DIBA and TAM in regulating recruitment of cofactors to chromatin (decreasing the interaction of ERα with AIB1 and blocking dissociation between ERα and NCoR caused by E2 or TAM). Consequently, DIBA restores the antagonistic action of TAM in breast cancer cells that have acquired resistance, in turn quenching target gene expression and blocking cell growth of TAM-resistant breast cancer cells. These studies suggest a possible new approach in modifying TAM resistance and a potential role for small electrophilic compounds that can modify the particularly vulnerable zinc finger in ERα.

Experimental procedures

Cell and cell culture

The electrophilic compound DIBA (NSC654077) was from the Laboratory of Cell Biology, National Cancer Institute. The human breast carcinoma cell lines MCF-7, ZR-75, and MDA-MB-468 were obtained from ATCC (Manassas, VA). The MCF-7/LCC2 cell line was from Dr. R. Clarke. MCF-7/HER2-18 and BT474 cell lines were from Dr. K. Osborne. 4-OH-TAM and 17β-Estrodial were purchased from Sigma-Aldrich (St Louis, MO). ICI 182780 was from Tocris (Ellisville, MO). In experiments with estrogen or TAM, cells were cultured in phenol red-free and DMEM or RPMI 1640 supplemented with 5% charcoal-dextran-stripped fetal calf serum for at least 2 days.

Cell proliferation and cell-cycle analysis

Cell proliferation was examined by measuring DNA synthesis with [^{3}H]thymidine uptake (Wang et al., 2004). Cell cycle was analyzed by propidium iodide staining and FACS (Li et al., 2006).

Electrophoretic mobility shift assay

Electrophoretic mobility shift assay (EMSA) was performed as described previously (Wang et al., 2003). End-labeled [^{32}P] oligonucleotide probes correspond to the ERE consensus sequence: 5′-GATCCGTCAGGTCAC AGTGACCTGATGGATC-3′, ARE consensus: 5′-GAAGTCTGGTACAGG GTGTTCTTTTTG-3′, and AP-1 consensus: 5′-CGCTTGATGAGTCAGCCG GAA-3′, respectively.

Expression plasmids

pSG5-HE0, pSG5-HE11, pSG5-HE16, or pSG5-HE19 expression plasmids were kindly provided by Dr. P. Chambon, Universitė Louis Pasteur, France. pSG5-HEZF was created by site-directed mutagenesis of HE0 using the oligonuecleotide: 3305: 5′-CTCACTATAGGGCGAATTCCGGCCACGGACCAT GACCATGACCC-3′; 3306: 5′-CATATAGTCGTTATGTCCTTGAATACTTCTC TTGAAGAAGGCCTTGTAGCGAGTCTCCTTGGCAGATTCC-3′; 3307: 5′-GG CCTTCTTCAAGAGAAGTATTCAAGGACATAACGACTATATGTACGAAGTGG GAATGATGAAAGGTGGG-3′; 3308: 5′-TCAGACTGTGGCAGGGAAACCC TCTGCCTCCCCC-3′; 3309: 5′-AACTCGAGCTGGATCCTCAGACTGTGGC AGGGAAACCCTCTGCCTCCCCC-3′ resulting in deletion of amino acids 185–205 and 221–245.

Transfection of SiRNA for ERα

BT474 cells were transfected with an ERα-SiRNA construct or control vector for 96 hr according to the manufacturer's instructions (New England BioLabs, MA). Efficacy of the constructs was tested through western blot analysis of the respective target ERα in transfected cells.

Transfection of luciferase reporter plasmids

FuGene-6 was used for transfection of luciferase reporter plasmids or cotransfection of reporter gene plasmids with ER expression plasmids. Luciferase assays were performed according to the manufacturer's instructions (BD PharMingen, San Diego, CA).

Coimmunoprecipitation and western blot analysis

Coimmunoprecipitation and western blots were performed as previously described (Yang et al., 2000). Antibodies against ER, phospho-ER, AIB1, phospho-AIB1, NCoR, HER2, phospho-HER2, AKT, phospho-AKT, MAPK, phospho-MAPK, phosphotyrosine, and phosphoserine were from Upstate Biotechnology (Lake Placid, NY).

Chromatin immunoprecipitation

The ChIP assays were based on a protocol described by Shang et al. (2000). Cells were fixed by formaldehyde. Purified chromatin samples were immunoprecipitated with anti-ERα antibody. DNA, isolated from immunoprecipitated material following reversal of formaldehyde crosslinking, was amplified by PCR. Promoter-specific primers included: pS2, 5′-CCGGCCATCTCTCAC TAT-3′ (forward primer) and 5′-ATCTTGGCTGAGGGATCT-3′ (reverse primer); pS2 upstream primer pair for negative control, 5′-GAAGACTCCG CACCTCAGAC-3′ (forward primer) and 5′-CCCTTGTGGGGAATCTGG-3′ (reverse primer); c-Myc, 5′-CCGCCTGCGATGATTTATAC-3′ (forward primer) and 5′-AAGGTGGGGAGGAGACTCAG-3′ (reverse primer); Cathepsin D, 5′-TCCAGACATCCTCTCTGGAA-3′ (forward primer), 5′-GGAGCGG AGGGTCCATTC-3′ (reverse primer). MMP-1 promote, 5′-TTGCAACACCAA GTGATTCCA-3′ (forward primer) and 5′-CCCAGCCTCTTGCTACTCCA-3′ (reverse primer); MMP-1 non-AP-1 specific site, 5′-GAGTACAACTTACA TCGTGTTGCAG-3′ (forward primer) and 5′-ATATGGCTTGGATGCCATCA ATGTC-3′ (forward primer).

ChIP Re-ChIP

Complexes were eluted from the primary immunoprecipitation by incubation with 10 mM DTT at 37°C for 30 min and diluted 1:50 in buffer (1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris-HCl [pH 8.1]), followed by reimmunoprecipitation with the second antibodies (Shang et al., 2000). ChIP Re-ChIPs of supernatants were done essentially as the primary IPs.

Live microscopy

MCF-7/LCC2 cells were grown on 14 mm coverslips in 35 mm plates and transfected with a pEGFP-C2-hERα construct using FuGene-6. Before ligand addition, the starved cells were pretreated with DIBA for 2 hr. Images were acquired at 10 min intervals with a Zeiss LSM 510 confocal microscope using a 40×/1.3 NA oil immersion objective lens (Stenoien et al., 2000).

Human tumor xenografts

Human MCF-7/LCC2-derived tumor xenografts were established in female athymic Ncr-nu/nu nude mice (National Cancer Institute, Frederick, MD) as described previously (Brunner et al., 1993; Wang et al., 2004). Tumor volume is calculated as $a^2 \times b \times 0.5$, where "a" is the width and "b" is the length of the tumor. Formalin-fixed tissue sections were embedded in paraffin, stained with hematoxylin and eosin, and examined under a light microscope. Animal experimentation was reviewed and approved by NCI's Animal Research Committee.

Acknowledgments

We are very grateful to Dr. A.T. Maynard for helpful discussion in the initial stage, Dr. P. Chambon for kindly providing ER expression plasmids, Drs. J. Hartley and D. Esposito for help making pSG5-HEZF constructs, Drs. M. R. Anver, S. Lawrence and K. Rogers for help in pathology, Dr. O. M. Z. Howard for help in animal experiments, Drs. K. Noer and W. Li for help in analysis of cell cycles, and Mr. B. Harris and E. Cho for help in image analysis. This research was supported by the Intramural Research Program of the Center for Cancer Research, NCI/NIH, and also funded in part with federal funds from the NCI under contract # NO1-CO-12400.

Received: December 18, 2005
Revised: April 8, 2006
Accepted: September 28, 2006
Published: December 11, 2006

References

Albanell, J., and Baselga, J. (2001). Unraveling resistance to Trastuzumab (Herceptin): IGF-I receptor, a new suspect. J. Natl. Cancer Inst. *93*, 1830–1832.

Anzick, S.L., Kononen, J., Walker, R.L., Azorsa, D.O., Tanner, M.M., Guan, X.Y., Sauter, G., Kallioniemi, O.P., Trent, J.M., and Meltzer, P.S. (1997). AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. Science *277*, 965–968.

Bain, D.L., Franden, M.A., McManaman, J.L., Takimoto, G.S., and Horwitz, K.B. (2000). The N-terminal region of the human progesterone A-receptor. Structure analysis and the influence of the DNA binding domain. J. Biol. Chem. *275*, 7313–7320.

Benz, C.C., Scott, G.K., Sarup, J.C., Johnson, R.M., Tripathy, D., Coronado, E., Shepard, H.M., and Osborne, C.K. (1993). Estrogen-dependent, tamoxifen-resistant tumorigenic growth of MCF-7 cells transfected with HER2/neu. Breast Cancer Res. Treat. *24*, 85–95.

Brockdorff, B.L., Heiberg, I., and Lykkesfeldt, A.E. (2003). Resistance to different antiestrogens is caused by different multi-factorial changes and is associated with reduced expression of IGF-1 receptor Ialpha. Endocr. Relat. Cancer *10*, 579–590.

Brodie, J., and McEwan, I.J. (2005). Intra-domain communication between the N-terminal and DNA-binding domains of the androgen receptor: modulation of androgen response element DNA binding. J. Mol. Endocrinol. *34*, 603–615.

Brunner, N., Frandsen, T.L., Holst-Hansen, C., Bei, M., Thompson, E.W., Wakeling, A.E., Lippman, M.E., and Clarke, R. (1993). MCF7/LCC2: a 4-hydroxytamoxifen-resistant human breast cancer variant that retains sensitivity to the steroidal antiestrogen ICI 182,780. Cancer Res. *53*, 3229–3232.

Brzozowski, A.M., Pike, A.C., Dauter, Z., Hubbard, R.E., Bonn, T., Engstrom, O., Ohman, L., Greene, G.L., Gustafsson, J.A., and Carlquist, M. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. Nature *389*, 753–758.

Campbell, R.A., Bhat-Nakshatri, P., Patel, N.M., Constantinidou, D., Ali, S., and Nakshatri, H. (2001). Phosphatidylinositol 3-kinase/AKT-mediated

activation of ERα: a new model for anti-estrogen resistance. J. Biol. Chem. *276*, 9817–9824.

DeNardo, D.G., Kim, H.T., Hilsenbeck, S., Cuba, V., Tsimelzon, A., and Brown, P.H. (2005). Global gene expression analysis of estrogen receptor transcription factor cross talk in breast cancer: identification of estrogen-induced/AP-1-dependent genes. Mol. Endocrinol. *19*, 362–378.

Font de Mora, J., and Brown, M. (2000). AIB1 is a conduit for kinase-mediated growth factor signaling to the estrogen receptor. Mol. Cell. Biol. *20*, 5041–5047.

Fujita, T., Kobayashi, Y., Wada, O., Tateishi, Y., Kitada, L., Yamamoto, Y., Takashima, H., Murayama, A., Yano, T., Baba, T., et al. (2003). Full activation of ERα activation function-1 induces proliferation of breast cancer cells. J. Biol. Chem. *278*, 26704–26714.

Gee, J.M., Robertson, J.F., Gutteridge, E., Ellis, I.O., Pinder, S.E., Rubini, M., and Nicholson, R.I. (2005). EGF/HER2/IGF receptor signalling and oestrogen receptor activity in clinical breast cancer. Endocr. Relat. Cancer *1*, S99–S111.

Glass, C.K., and Rosenfeld, M.G. (2000). The coregulator exchange in transcriptional functions of nuclear receptors. Genes Dev. *14*, 121–141.

Graham, J.D., Bain, D.L., Richer, J.K., Jackson, T.A., Tung, L., and Horwitz, K.B. (2000). Nuclear receptor conformation, coregulators, and tamoxifen-resistant breast cancer. Steroids *65*, 579–584.

Gutierrez, M.C., Detre, S., Johnston, S., Mohsin, S.K., Shou, J., Allred, D.C., Schiff, R., Osborne, C.K., and Dowsett, M. (2005). Molecular changes in tamoxifen-resistant breast cancer: relationship between ER, HER-2, and p38 mitogen-activated protein kinase. J. Clin. Oncol. *23*, 2469–2476.

Hoffmann, J., Bohlmann, R., Heinrich, N., Hofmeister, H., Kroll, J., Kunzer, H., Lichtner, R.B., Nishino, Y., Parczyk, K., Sauer, G., et al. (2004). Characterization of new ER destabilizing compounds: effects on estrogen-sensitive and tamoxifen-resistant breast cancer. J. Natl. Cancer Inst. *96*, 210–218.

Ibrahim, Y.H., and Yee, D. (2005). Insulin-like growth factor-I and breast cancer therapy. Clin. Cancer Res. *11*, 944S–950S.

Jakacka, M., Ito, M., Weiss, J., Chien, P.Y., Gehm, B.D., and Jameson, J.L. (2001). Estrogen receptor binding to DNA is not required for its activity through the nonclassical AP1 pathway. J. Biol. Chem. *276*, 13615–13621.

Joel, P.B., Smith, J., Sturgill, T.W., Fisher, T.L., Blenis, J., and Lannigan, D.A. (1998). pp90rsk1 regulates ER-mediated transcription through phosphorylation of Ser-167. Mol. Cell. Biol. *18*, 1978–1984.

Jordan, V.C. (2004). Selective estrogen receptor modulation: concept and consequences in cancer. Cancer Cell *5*, 207–213.

Keeton, E.K., and Brown, M. (2005). Cell cycle progression stimulated by tamoxifen-bound estrogen receptor α and promoter-specific effects in breast cancer cells deficient in NCoR and SMRT. Mol. Endocrinol. *19*, 1543–1554.

Klinge, C.M., Jernigan, S.C., Mattingly, K.A., Risinger, K.E., and Zhang, J. (2004). Estrogen response element-dependent regulation of transcriptional activation of estrogen receptors alpha and beta by coactivators and corepressors. J. Mol. Endocrinol. *33*, 387–410.

Konecny, G., Pauletti, G., Pegram, M., Untch, M., Dandekar, S., Aguilar, Z., Wilson, C., Rong, H.M., Bauerfeind, I., Felber, M., et al. (2003). Quantitative association between HER-2/neu and steroid hormone receptors in hormone receptor-positive primary breast cancer. J. Natl. Cancer Inst. *5*, 142–153.

Kumar, R., and Thompson, E.B. (2003). Transactivation functions of the N-terminal domains of nuclear hormone receptors: protein folding and coactivator interactions. Mol. Endocrinol. *17*, 1–10.

Kumar, V., Green, S., Stack, G., Berry, M., Jin, J.R., and Chambon, P. (1987). Functional domains of the human estrogen receptor. Cell *51*, 941–951.

Kumar, R., Baskakov, I.V., Srinivasan, G., Bolen, D.W., Lee, J.C., and Thompson, E.B. (1999). Interdomain signaling in a two-domain fragment of the human glucocorticoid receptor. J. Biol. Chem. *274*, 24737–24741.

Kurebayashi, J., Otsuki, T., Kunisue, H., Tanaka, K., Yamamoto, S., and Sonoo, H. (2000). Expression levels of estrogen receptor-α, estrogen receptor-β, coactivators, and corepressors in breast cancer. Clin. Cancer Res. *6*, 512–518.

Kushner, P.J., Agard, D.A., Greene, G.L., Scanlan, T.S., Shiau, A.K., Uht, R.M., and Webb, P. (2000). Estrogen receptor pathways to AP-1. J. Steroid Biochem. Mol. Biol. *74*, 311–317.

Laity, J.H., Lee, B.M., and Wright, P.E. (2001). Zinc finger proteins: new insights into structural and functional diversity. Curr. Opin. Struct. Biol. *11*, 39–46.

Lannigan, D.A. (2003). Estrogen receptor phosphorylation. Steroids *68*, 1–9.

Lavinsky, R.M., Jepsen, K., Heinzel, T., Torchia, J., Mullen, T.M., Schiff, R., Del-Rio, A.L., Ricote, M., Ngo, S., Gemsch, J., et al. (1998). Diverse signaling pathways modulate nuclear receptor recruitment of NCoR and SMRT complexes. Proc. Natl. Acad. Sci. USA *95*, 2920–2925.

Li, W.Q., Jiang, Q., Aleem, E., Kaldis, P., Khaled, A.R., and Durum, S.K. (2006). IL-7 promotes T-cell proliferation through destabilization of p27Kip1. J. Exp. Med. *203*, 573–582.

Lilling, G., Hacohen, H., Nordenberg, J., Livnat, T., Rotter, V., and Sidi, Y. (2000). Differential sensitivity of MCF-7 and LCC2 cells, to multiple growth inhibitory agents: possible relation to high bcl-2/bax ratio? Cancer Lett. *8*, 27–34.

Lin, Y.Z., Li, S.W., and Clinton, G.M. (1990). Insulin and epidermal growth factor stimulate phosphorylation of p185HER-2 in the breast carcinoma cell line, BT474. Mol. Cell. Endocrinol. *69*, 111–119.

Mak, H.Y., Hoare, S., Henttu, P.M.S., and Parker, M.G. (1999). Molecular determinants of the estrogen receptor-coactivator interface. Mol. Cell. Endocrinol. *19*, 3895–3903.

Martin, M.B., Franke, T.F., Stoica, G.E., Chambon, P., Katzenellenbogen, B.S., Stoica, B.A., McLemore, M.S., Olivo, S.E., and Stoica, A. (2000). A role for Akt in mediating the estrogenic functions of EGF and IGF-1. Endocrinology *141*, 4503–4511.

Maynard, A.T., and Covell, D.G. (2001). Reactivity of zinc finger cores: analysis of protein packing and electrostatic screening. J. Am. Chem. Soc. *123*, 1047–1058.

McDonnell, D.P., and Norris, J.D. (2002). Connections and regulation of the human estrogen receptor. Science *296*, 1642–1644.

McKenna, N.J., and O'Malley, B.W. (2002). Minireview: nuclear receptor coactivators—an update. Endocrinology *143*, 2461–2465.

Osborne, C.K. (1998). Tamoxifen in the treatment of breast cancer. N. Engl. J. Med. *339*, 1609–1618.

Osborne, C.K., Bardou, V., Hopp, T.A., Chamness, G.C., Hilsenbeck, S.G., Fuqua, S.A.W., Wong, J., Allred, D.C., Clark, G.M., and Schiff, R. (2003). Role of the estrogen receptor coactivator AIB1 (SRC-3) and HER-2/neu in tamoxifen resistance in breast cancer. J. Natl. Cancer Inst. *95*, 353–361.

Osborne, C.K., Shou, J., Massarweh, S., and Schiff, R. (2005). Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer. Clin. Cancer Res. *11*, 865S–870S.

Predki, P.F., and Sarkar, B. (1992). Effect of replacement of ''zinc finger'' zinc on estrogen receptor DNA interactions. J. Biol. Chem. *267*, 5842–5846.

Ruff, M., Gangloff, M., Wurtz, J.M., and Moras, D. (2000). Estrogen receptor transcription and transactivation: structure-function relationship in DNA- and ligand-binding domains of estrogen receptors. Breast Cancer Res. *2*, 353–359.

Schiff, R., Massarweh, S.A., Shou, J., Bharwani, L., Mohsin, S.K., and Osborne, C.K. (2004). Cross-talk between ER and growth factor pathways as a molecular target for overcoming endocrine resistance. Clin. Cancer Res. *10*, 331S–336S.

Schoenmakers, E., Alen, P., Verrijdt, G., Peeters, B., Verhoeven, G., Rombauts, W., and Claessens, F. (1999). Differential DNA binding by the androgen and glucocorticoid receptors involves the second Zn-finger and a C-terminal extension of the DNA-binding domains. Biochem. J. *341*, 515–521.

Schwabe, J.W.R., Chapman, L., Finch, J.T., and Rhodes, D. (1993). The crystal structure of the estrogen receptor DNA-binding domain bound to

DNA: how receptors discriminate between their response elements. Cell *75*, 567–578.

Shang, Y., and Brown, M. (2002). Molecular determinants for the tissue specificity of SERMs. Science *295*, 2465–2468.

Shang, Y., Hu, X., DiRenzo, J., Lazar, M.A., and Brown, M. (2000). Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. Cell *103*, 843–852.

Shao, D., Rangwala, S.M., Bailey, S.T., Krakow, S.L., Reginato, M.J., and Lazar, M.A. (1998). Interdomain communication regulating ligand binding by PPAR-γ. Nature *396*, 377–380.

Shiau, A.K., Barstad, D., Loria, P.M., Cheng, L., Kushner, P.J., Agard, D.A., and Greene, G.L. (1998). The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell *95*, 927–937.

Shou, J., Massarweh, S., Osborne, C.K., Wakeling, A.E., Ali, S., Weiss, H., and Schiff, R. (2004). Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/neu cross-talk in ER/HER2-positive breast cancer. J. Natl. Cancer Inst. *96*, 926–935.

Smith, C.L., Nawaz, Z., and O'Malley, B.W. (1997). Coactivator and corepressor regulation of the agonist/antagonist activity of the mixed antiestrogen, 4-hydroxytamoxifen. Mol. Endocrinol. *11*, 657–666.

Stenoien, D.L., Mancini, M.G., Patel, K., Allegretto, E.A., Smith, C.L., and Mancini, M.A. (2000). Subnuclear trafficking of ERα and steroid receptor coactivator-1. Mol. Endocrinol. *14*, 518–534.

Stoica, A., Saceda, M., Doraiswamy, V.L., Coleman, C., and Martin, M.B. (2000). Regulation of ERα gene expression by EGF. J. Endocrinol. *165*, 371–378.

Takimoto, G.S., Tung, L., Abdel-Hafiz, H., Abel, M.G., Sartorius, C.A., Richer, J.K., Jacobsen, B.M., Bain, D.L., and Horwitz, K.B. (2003). Functional properties of the N-terminal region of progesterone receptors and their mechanistic relationship to structure. J. Steroid Biochem. Mol. Biol. *85*, 209–219.

Tikkanen, M.K., Carter, D.J., Harris, A.M., Le, H.M., Azorsa, D.O., Meltzer, P.S., and Murdoch, F.E. (2000). Endogenously expressed ER and coactivator AIB1 interact in MCF-7 human breast cancer cells. Proc. Natl. Acad. Sci. USA *97*, 12536–12540.

Tsai, M.J., and O'Malley, B.W. (1994). Molecular mechanisms of action of steroid/thyroid receptor superfamily members. Annu. Rev. Biochem. *63*, 451–486.

Voss, T.C., Demarco, I.A., Booker, C.F., and Day, R.N. (2005). Corepressor subnuclear organization is regulated by ER via a mechanism that requires the DNA-binding domain. Mol. Cell. Endocrinol. *231*, 33–47.

Wang, L.H., Yang, X.Y., Zhang, X.H., Mihalic, K., Xiao, W., and Farrar, W.L. (2003). The *cis* decoy against the estrogen-responsive element inhibits breast cancer cell proliferation via target suppressing c-fos not mitogen-activated protein kinase activity. Cancer Res. *63*, 2046–2051.

Wang, L.H., Yang, X.Y., Mihalic, K., Zhang, X.H., Xiao, W., Maynard, A.T., and Farrar, W.L. (2004). Suppression of breast carcinoma by chemical modulation of the vulnerable zinc finger in estrogen receptor. Nat. Med. *10*, 40–47.

Webb, P., Nguyen, P., Valentine, C., Lopez, G.N., Kwok, G.R., McInerney, E., Katzenellenbogen, B.S., Enmark, E., Gustafsson, J.-Å., Nilsson, S., et al. (1999). The estrogen receptor enhances AP-1 activity by two distinct mechanisms with different requirements for receptor transactivation functions. Mol. Endocrinol. *13*, 1672–1685.

Wikstrom, A., Berglund, H., Hambraeus, C., van den Berg, S., and Hard, T. (1999). Conformational dynamics and molecular recognition: backbone dynamics of the estrogen receptor DNA-binding domain. J. Mol. Biol. *289*, 963–979.

Yamashita, H., Nishio, M., Kobayashi, S., Ando, Y., Sugiura, H., Zhang, Z., Hamaguchi, M., Mita, K., Fujii, Y., and Iwase, H. (2005). Phosphorylation of ERα serine 167 is predictive of response to endocrine therapy and increases postrelapse survival in metastatic breast cancer. Breast Cancer Res. *7*, R753–R764.

Yang, X.Y., Wang, L.H., Chen, T., Hodge, D.R., Resau, J.H., DaSilva, L., and Farrar, W.L. (2000). Activation of human T lymphocytes is inhibited by PPARγ agonists. PPARγ co-association with transcription factor NFAT. J. Biol. Chem. *275*, 4541–4544.

ELSEVIER

Cancer Genetics and Cytogenetics ■ (2007) ■

Cancer Genetics and Cytogenetics

Short communication

The A4396G polymorphism in interferon regulatory factor 1 is frequently expressed in breast cancer cell lines

Kerrie B. Bouker, Todd C. Skaar[1], David S. Harburger, Rebecca B. Riggins, David R. Fernandez, Alan Zwart, Robert Clarke*

Lombardi Comprehensive Cancer Center and Department of Oncology, Georgetown University School of Medicine, Room W405A Research Building, 3970 Reservoir Road NW, Washington, DC 20057

Received 7 November 2006; received in revised form 20 December 2006; accepted 22 December 2006

Abstract Loss or mutation of known tumor suppressor genes accounts for a small proportion of all breast cancers. We have recently shown that interferon regulatory factor 1 (*IRF1*) is a putative tumor suppressor gene in breast cancer. We now report that the A4396G single nucleotide polymorphism in the *IRF1* gene is more frequent in human breast cancer cell lines than in the general population ($P = 0.01$). Furthermore, A4396G is more frequently expressed in African American (black) than in European ancestry (white) subjects ($n = 70$ subjects; $P \leq 0.001$), leading to a significant difference in genotype distribution between these populations ($P = 0.002$).

1. Introduction

The precise molecular events responsible for affecting breast cancer risk and disease progression remain to be established. Although an increasing number of oncogenes have been identified, relatively few tumor suppressor genes have been implicated in driving the development and progression of this disease. The transcription factor interferon regulatory factor 1 gene (HUGO symbol *IRF1*; http://www.gene.ucl.ac.uk/nomenclature) is lost, mutated, or rearranged in several cancers including some hematopoietic [1] and gastric cancers [2]. Several single nucleotide polymorphisms have also been reported [3,4]. We have previously shown that *IRF1* is associated with acquired estrogen-independence [5] and with antiestrogen resistance in breast cancer [6]; others have shown its importance in normal mammary epithelial cells [7] and in other breast cancer models [8]. The *IRF1* protein is readily detected in breast tumors [9,10] and is expressed in a pattern consistent with a putative gene network that regulates endocrine responsiveness in breast cancer cells [10,11].

We have shown that *IRF1* acts as a breast cancer suppressor gene, repressing both the growth of human breast cancer cell xenografts in athymic nude mice and cell proliferation in vitro [12]. These tumor suppressor and antiproliferative activities are associated with the ability of *IRF1* to modulate apoptosis through regulation of a caspase cascade [7,12]. Loss of heterozygosity at 5q31.1, which includes the *IRF1* gene locus, has recently been reported in both sporadic and inherited breast cancers [13,14].

We now report that a polymorphism in the *IRF1* gene, an A→G single nucleotide polymorphism at base pair 4396 (A4396G), is detected at a higher frequency in human breast cancer cell lines than in the normal population and is more frequently expressed in African Americans than in European-ancestry whites.

2. Materials and methods

MCF-7 cells were originally obtained from Dr. Marvin Rich (Michigan Cancer Foundation, Detroit, MI). T47D, ZR-75-1, MDA-MB-231, MDA-MB-435, and A1N4 cells were obtained from the Tissue Culture Shared Resource of the Georgetown University Lombardi Comprehensive Cancer Center. All other cell lines were obtained from the American Type Culture Collection (ATCC, Manassas, VA). Unless otherwise indicated, cell lines were routinely grown in improved minimal essential medium (IMEM; Biofluids, Rockville, MD) with phenol red and supplemented with 5% fetal bovine serum (Gibco Life Technologies—Invitrogen, Carlsbad, CA). A1N4 cells were grown in

* Corresponding author. Tel.: (202) 687-3755; fax: (202) 687-7505.
E-mail address: clarker@georgetown.edu (R. Clarke).
[1] Present address: Department of Medicine, Division of Clinical Pharmacology and Indiana University Cancer Center, Indiana University, Indianapolis, IN 46202.

0165-4608/07/$ − see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cancergencyto.2006.12.008

IMEM supplemented with 0.5% fetal calf serum, 0.5 μg/mL hydrocortisone, 5 μg/mL insulin, and 10 ng/mL epidermal growth factor [15]. All cells were maintained in a humidified incubator at 37°C in an atmosphere containing 95% air, 5% CO_2.

Restriction fragment length polymorphism (RFLP) analysis was performed on a fragment of the *IRF1* gene amplified from genomic DNA by polymerase chain reaction (PCR). For DNA extraction, cells were plated in T-75cm^2 plastic tissue culture flasks at a density of 1×10^6 cells/flask and grown for 24 hours prior to DNA isolation. DNA was extracted from proliferating subconfluent cells using the TRIzol reagent (~~Life Technologies, Gaithersburg, MD~~) and quantified by comparing the optical density ratios (OD_{260}/OD_{280}) obtained spectrophotometrically using a Beckman DU640 spectrophotometer (Beckman, Fullerton, CA). For RFLP analysis, *IRF1* specific primers were used to amplify a portion of exon 7 in the *IRF1* gene from genomic DNA. The *IRF1* primers were bp-4358 (sense), 5′-TTGACCTGTGGCTTCTGCTGT-3′, and bp-4639 (antisense), 5′-GTCGCTTGCCTCCCCCTATG-3′ from GenBank (accession no. L05072; http://www.ncbi.nlm.nih.gov); 100 ng of the appropriate genomic DNA was used as template. PCR conditions were 1 cycle for 3 min at 94°C, 35 cycles of 1 min at 94°C and 1 min at 59°C, and 1 min 45 sec at 72°C. Purified PCR product was digested with the restriction endonuclease *Nla*III (New England BioLabs, Beverly, MA), size fractionated on a 20% Tris–boric acid–EDTA polyacrylamide gel, and stained with ethidium bromide. *Nla*III is site-specific (recognition sequence: C<u>A</u>TG) and will not cleave variant DNA at the 4396 base pair. Thus, two bands correspond to a homozygote A→G polymorphism at 4396, three bands signify a homozygous wild type, and four bands indicate a heterozygote.

For statistical analyses, the proportions of genotypes and allele frequencies in cell lines and human subjects were compared by χ^2 analysis. Unless otherwise indicated, all probabilities are two-tailed; the conventional assessment of $P < 0.05$ was applied to establish statistical significance.

3. Results and discussion

IRF1 is implicated as a tumor suppressor gene in breast cancer [12] and in several hematopoietic [1] and gastric cancers [2]. We found the A4396G polymorphism in the *IRF1* gene when sequencing PCR products from MCF-7 human breast cancer cells. Using RFLP analysis, we measured the prevalence of the polymorphism in breast cancer cell lines (Table 1; $n = 17$), other cancer cell lines (Table 2; $n = 40$), normal cell lines (Tables 1 and 2; $n = 5$), and in DNA from normal volunteers obtained from the Coriell DNA repository (Coriell Institute for Medical Resarch, Camden, NJ): 34 African Americans and 36 whites ($n = 70$) (Table 3). Allele frequencies are significantly different between African Americans (χ^2 test, $P = 0.004$; higher frequency of G vs. A) and whites (χ^2 test, $P < 0.001$; higher frequency of A vs. G). The distribution of genotypes (allelic frequency of A was 0.37 in African Americans and 0.65 in whites) was significantly different between African Americans and whites (χ^2 test, $P = 0.002$).

Table 1
A4396G polymorphism in breast cancer and normal mammary epithelial cell lines

Cell line	bp 4396	Ethnicity[a]
MCF7	G/G	Eur
MCF7	A/G	Eur
NIH/ADR-RES	A/A	unknown
ZR-75-1	A/A	Eur
ML-20 (MCF-7 transfected with syk)	A/G	Eur
MKL-4 (MCF-7 transfected with FGF4)	A/G	Eur
BRC-230	A/A	unknown
BT-474	A/G	Eur
BT-483	A/A	Eur
BT-549	G/G	Eur
DU4475	A/A	Eur
Hs 578T	A/A	Eur
MDA-MB-134-VI	A/A	Eur
MDA-MB-157	G/G	AA
MDA-MB-231	G/G	Eur
MDA-MB-415	A/A	Eur
MDA-MB-330	G/G	Eur
MDA-MB-453	A/A	Eur
MDA435/LCC6	G/G	unknown
Normal mammary cell lines		
A1N4	A/A	
MCF10A	A/A	

Abbreviations: AA, African American (black); Eur, European ancestry (white).

[a] Ethnicity data were obtained from the original publications or as described by ATCC.

We compared allele frequencies among the breast cancer cell lines for which we could identify ethnic origin as white from either the ATCC (Manassas, VA) or the original publications (Table 1) with the white population data from Table 3. We excluded MDA-MB-157 cells obtained from an African American woman, the MCF-7 cells because we found two different genotypes, and NIH/ADR-RES cells because the precise origin of these cells is uncertain [16]. We excluded MDA435/LCC6 cells [17]; the breast origin of the parental MDA-MB-435 cell line has been questioned [18], and in a recent study we could not show similarity in the transcriptomes of MDA-MB-435 cells and a series of breast cancer biopsies, whereas other breast cancer cell lines show significant similarities with these breast tumors [19]. These cells now appear to be definitively of melanoma origin [20]. We also excluded two normal mammary epithelial cell lines, A1N4 [21] and MCF10A [22]. The remaining breast cancer cell lines ($n = 11$) have a genotype proportion significantly different from that of the white population (χ^2 test; $P = 0.01$); genotype frequencies in the breast cancer cell lines are not

Table 2
A4396G polymorphism in cancer and other cell lines not of mammary origin

Cell line	Tissue of origin	bp 4396
CCF STTG1	Astrocytoma	A/A
HeLa	Cervical adenocarcinoma	A/G
LS 147T	Colorectal adenocarcinoma	A/G
LS 180	Colon adenocarcinoma	A/G
CaCo-2	Colon adenocarcinoma	A/A
HCT-116	Colon adenocarcinoma	A/A
HCT-15	Colorectal adenocarcinoma	G/G
HEC-1-A	Endometrial adenocarcinoma	A/A
A-431	Epidermoid squamous carcinoma	A/A
Hs 913T	Fibrosarcoma	A/G
KATO III	Gastric adenocarcinoma	A/G
A-172	Glioblastoma	A/A
HepG2	Hepatocellular carcinoma	A/A
K-562	Leukemia (chronic myeloid blast crisis)	A/A
Jurkat	Leukemia (T-Lymphocyte)	G/G
IMR-90	Lung fibroblasts (normal)	A/G
A-549	Lung (bronchoalveolar carcinoma)	A/A
Calu-3	Lung adenocarcinoma	A/A
NCI-H209	Lung adenocarcinoma	A/A
NCI-H345	Lung carcinoma	A/A
NCI-H520	Lung squamous cells carcinoma	G/G
MOLT-4	Lymphoblastic leukemia (T-cell)	A/A
MOLT-3	Lymphoblastic leukemia (T-cell)	A/A
CCRF-CEM	Lymphoblastic leukemia (T-cell)	A/A
CCRF-HSB-2	Lymphoblastic leukemia (T-cell)	A/A
CCRF-SB	Lymphoblastic leukemia (T-cell)	A/A
Daudi	Lymphoma (Burkitt)	G/G
NHL	Lymphoma (Non-Hodgkin)	A/G
IMR-32	Neuroblastoma	A/G
BE (2) M17	Neuroblastoma	G/G
Hs 683	Neuroglioma	A/G
CaOV-3	Ovarian adenocarcinoma	G/G
COLO-357	Pancreatic adenocarcinoma	G/G
FaDu	Pharyngeal carcinoma	A/A
BeWo	Choriocarcinoma	A/G
JAR	Choriocarcinoma	A/G
JEG-3	Choriocarcinoma	A/G
HL-60	Promyelocytic leukemia	A/A
LNCaP	Prostate adenocarcinoma	G/G
DU 145	Prostate carcinoma	A/A
A-204	Rhabdomyosarcoma	A/A
A-673	Rhabdomyosarcoma	G/G
Hs-27	Foreskin fibroblast (normal)	A/A
Hs-68	Foreskin fibroblast (normal)	A/G

significantly different from those in a random sample of other cancer cell lines (Table 2; $n = 40$). Of the excluded breast cancer cell lines, only the NIH/ADR-RES genotype is wild type, and the outcome is unaffected when all the cancer cell lines from Table 2 are compared.

Table 3
A4396G polymorphism in healthy human volunteers

Ethnicity	A/A	A/G	G/G
African American ($n = 34$)	3	19	12
European ancestry ($n = 36$)	14	19	3

The greater prevalence of A4396G in breast cancer cell lines derived from whites, compared with that in the normal white population, suggests an association with breast cancer. The genotype proportions in breast cancer cells are also seen in other cancer cell lines, further suggesting an association of A4396G with cancer. Nonetheless, we cannot fully exclude the possibility that this is a cell culture artifact. Our identification of two MCF-7 genotypes (A/G; G/G) suggests genetic drift in vitro, which may contribute to the phenotypic diversity of this cell line. Ongoing studies are now examining the prevalence of this polymorphism in breast tumors.

How the A4396G polymorphism contributes to the tumor suppressor role of *IRF1* in breast cancer is unknown. Noguchi et al. [4] first identified the A4396G polymorphism in peripheral blood lymphocytes from several atopic and asthmatic families of Japanese descent, although neither allele was significantly associated with transmission to asthmatic children. The switch in nucleotide use does not change the amino acid sequence of the translated protein. Seven splice variants have been previously reported in the *IRF1* gene [23]; alternative splicing in exons 7, 8, and 9 negatively regulates *IRF1* in cervical cancer, and this splicing likely affects its tumor suppressor activities [24]. It is not known if A4396G is an active splice site but in silico analysis using GeneSplicer [25] suggests that this polymorphism is unlikely to generate a novel splice site. In contrast, the A4396G polymorphism may affect putative transcription factor binding at internal sites [26]. Compared with the wild type allele, A4396G loses binding sites for microphthalmia transcription factor (*MITF/TFE3*), paraxis (TCF15), neurogenin 1 and 3, and Myc-Max heterodimers; these sites are replaced with a single hairy and enhancer of split 1 (*HES1*) site. In melanoma, MITF transcriptionally activates the tumor suppressor and cell cycle inhibitor INK4A, leading to cell cycle arrest [27]. Whether MITF or any of these other transcription factors differentially regulates the wild type versus A4396G *IRF1* allele is under investigation, because altered regulation of *IRF1* tumor suppressor activities could significantly affect breast cancer risk.

Of interest is the significantly higher prevalence of A4396G in African Americans. African American women are diagnosed at an earlier age [28] and present with a higher stage at diagnosis [29]. Although the incidence of breast cancer is lower [30], except for very young women [29], survival also is lower for African American than for non-Hispanic white and Hispanic women [31]. The increased prevalence of the A4396G polymorphism, particularly if it affects *IRF1*-mediated tumor suppression, could contribute to these observations in African American women, but this remains to be established.

References

[1] Willman CL, Sever CE, Pallavicini MG, Harada H, Tanaka N, Slovak ML, Yamamoto H, Harada K, Meeker TC, List AF, Taniguchi T. Deletion of IRF-1, mapping to chromosome 5q31.1,

in human leukemia and preleukemic myelodysplasia. Science 1993;259:965−71.

[2] Nozawa H, Oda E, Ueda S, Tamura G, Maesawa C, Muto T, Taniguchi T, Tanaka N. Functionally inactivating point mutation in the tumor-suppressor IRF-1 gene identified in human gastric cancer. Int J Cancer 1998;77:522−7.

[3] Ji H, Ball TB, Kimani J, Plummer FA. Novel interferon regulatory factor-1 polymorphisms in a Kenyan population revealed by complete gene sequencing. J Hum Genet 2004;49:528−35.

[4] Noguchi E, Shibasaki M, Arinami T, Yamakawa-Kobayashi K, Yokouchi Y, Takeda K, Matsui A, Hamaguchi H. Mutation screening of interferon regulatory factor 1 gene (IRF-1) as a candidate gene for atopy/asthma. Clin Exp Allergy 2000;30:1562−7.

[5] Gu Z, Lee RY, Skaar TC, Bouker KB, Welch JN, Lu J, Liu A, Zhu Y, Davis N, Leonessa F, Brunner N, Wang Y, Clarke R. Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-κB, and cyclic AMP response element binding with acquired resistance to faslodex (ICI 182,780). Cancer Res 2002;62:3428−37.

[6] Bouker KB, Skaar TC, Fernandez DR, O'Brien KA, Clarke R. Interferon regulatory factor-1 mediates the proapoptotic but not cell cycle arrest effects of the steroidal antiestrogen ICI 182,780 (Faslodex, Fulvestrant). Cancer Res 2004;64:4030−9.

[7] Bowie ML, Dietze EC, Delrow J, Bean GR, Troch MM, Marjoram RJ, Seewaldt VL. Interferon-regulatory factor-1 is critical for tamoxifen-mediated apoptosis in human mammary epithelial cells. Oncogene 2004;23:8743−55.

[8] Pizzoferrato E, Liu Y, Gambotto A, Armstrong MJ, Stang MT, Gooding WE, Alber SM, Shand SH, Watkins SC, Storkus WJ, Yim JH. Ectopic expression of interferon regulatory factor-1 promotes human breast cancer cell death and results in reduced expression of survivin. Cancer Res 2004;64:8381−8.

[9] Doherty GM, Boucher L, Sorenson K, Lowney J. Interferon regulatory factor expression in human breast cancer. Ann Surg 2001;233: 623−9.

[10] Zhu Y, Singh B, Hewitt S, Liu A, Gomez B, Wang A, Clarke R. Expression patterns among interferon regulatory factor-1, human X-box binding protein-1, nuclear factor kappa B, nucleophosmin, estrogen receptor alpha and progesterone receptor proteins in breast cancer tissue microarrays. Int J Oncol 2006;28:67−76.

[11] Clarke R, Liu MC, Bouker KB, Gu Z, Lee RY, Zhu Y, Skaar TC, Gomez B, O'Brien K, Wang Y, Hilakivi-Clarke LA. Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling. Oncogene 2003;22:7316−39.

[12] Bouker KB, Skaar TC, Riggins R, Harburger DS, Fernandez DR, Zwart A, Wang A, Clarke R. Interferon regulatory factor-1 (IRF-1) exhibits tumor suppressor activities in breast cancer associated with caspase activation and induction of apoptosis. Carcinogenesis 2005;26:1527−35.

[13] Johannsdottir HK, Jonsson G, Johannesdottir G, Agnarsson BA, Eerola H, Arason A, Heikkila P, Egilsson V, Olsson H, Johannsson OT, Nevanlinna H, Borg A, Barkardottir RB. Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors. Int J Cancer 2006;119:1052−60.

[14] Loo LW, Grove DI, Williams EM, Neal CL, Cousens LA, Schubert EL, Holcomb IN, Massa HF, Glogovac J, Li CI, Malone KE, Daling JR, Delrow JJ, Trask BJ, Hsu L, Porter PL. Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. Cancer Res 2004;64:8541−9.

[15] Stampfer MR, Bartley JC. Human mammary epithelial cells in culture: differentiation and transformation. In: Lippman ME, Dickson RB, editors. Breast cancer: cellular and molecular biology. Boston: Kluwer Academic Publishers, 1988. pp. 1−24.

[16] Scudiero DA, Monks A, Sausville EA. Cell line designation change: multidrug-resistant cell line in the NCI anticancer screen. J Natl Cancer Inst 1998;90:862−3.

[17] Leonessa F, Green D, Licht T, Wright A, Wingate-Legette K, Lippman J, Gottesman MM, Clarke R. MDA435/LCC6 and MDA435/LCC6^{MDR1}: ascites models of human breast cancer. Br J Cancer 1996;73:154−61.

[18] Rae JM, Ramus SJ, Waltham M, Armes JE, Campbell IG, Clarke R, Barndt RJ, Johnson MD, Thompson EW. Common origins of MDA-MB-435 cells from various sources with those shown to have melanoma properties. Clin Exp Metastasis 2004;21:543−52.

[19] Zhu Y, Wang A, Liu MC, Zwart A, Lee RY, Gallagher A, Wang Y, Miller WR, Dixon JM, Clarke R. Estrogen receptor alpha (ER) positive breast tumors and breast cancer cell lines share similarities in their transcriptome data structures. Int J Oncol 2006;28:67−76.

[20] Rae JM, Creighton CJ, Meck JM, Haddad BR, Johnson MD. MDA-MB-435 cells are derived from M14 Melanoma cells: a loss for breast cancer, but a boon for melanoma research. Breast Cancer Res Treat 2006.

[21] Valverius EM, Walker-Jones D, Bates SE, Stampfer M, Clark R, McCormick F, Dickson RB, Lippman ME. Production of and responsiveness to transforming growth factor beta in normal and oncogene-transformed human mammary epithelial cells. Cancer Res 1989;49: 6269−74.

[22] Tait L, Soule H, Russo J. Ultrastructural and immunocytochemical characterization of a immortalized human breast epithelial cell line, MCF-10. Cancer Res 1990;50:6087−94.

[23] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel), 1997. GeneCard for IRF-1 [Internet]. Available at: http://www.genecards.org/cgi-bin/carddisp.pl?gene = IRF1 Updated Oct. 29, 2006. [Q3][Q4]

[24] Lee EJ, Jo M, Park J, Zhang W, Lee JH. Alternative splicing variants of IRF-1 lacking exons 7, 8, and 9 in cervical cancer. Biochem Biophys Res Commun 2006;347:882−8.

[25] Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 2001;29: 1185−90.

[26] Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 2005;21:2933−42.

[27] Loercher AE, Tank EM, Delston RB, Harbour JW. MITF links differentiation with cell cycle arrest in melanocytes by transcriptional activation of INK4A. J Cell Biol 2005;168:35−40.

[28] Aziz H, Hussain F, Sohn C, Mediavillo R, Saitta A, Hussain A, Brandys M, Homel P, Rotman M. Early onset of breast carcinoma in African American women with poor prognostic factors. Am J Clin Oncol 1999;22:436−40.

[29] Weir HK, Thun MJ, Hankey BF, Ries LA, Howe HL, Wingo PA, Jemal A, Ward E, Anderson RN, Edwards BK. Annual report to the nation on the status of cancer, 1975−2000, featuring the uses of surveillance data for cancer prevention and control. J Natl Cancer Inst 2003;95:1276−99.

[30] Ghafoor A, Jemal A, Ward E, Cokkinides V, Smith R, Thun M. Trends in breast cancer by race and ethnicity. CA Cancer J Clin 2003;53:342−55. [Erratum in: CA Cancer J Clin 2004;54:181].

[31] Shavers VL, Harlan LC, Stevens JL. Racial/ethnic variation in clinical presentation, treatment, and survival among breast cancer patients under age 35. Cancer 2003;97:134−47.

BMC Cancer

BioMed Central

Research article

AIB1 gene amplification and the instability of polyQ encoding sequence in breast cancer cell lines

Lee-Jun C Wong*[1], Pu Dai[2], Jyh-Feng Lu[3], Mary Ann Lou[4], Robert Clarke[5] and Viktor Nazarov[5]

Address: [1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA, [2]Department of Otolaryngology, Head Neck Surgery, Chinese PLA General Hospital, Beijing 100853, China, [3]Fu Jen Catholic University, School of Medicine, Taipei, Taiwan, [4]Department of Surgery, Cardinal Tien Hospital, Hsintien Taipei Hsien, Taiwan and [5]Department of Oncology, Georgetown University Medical Center, Washington, DC 20007, USA

Email: Lee-Jun C Wong* - ljwong@bcm.edu; Pu Dai - daipu301@yahoo.com; Jyh-Feng Lu - med0001@mails.fju.edu.tw; Mary Ann Lou - tienhop8@ms55.hinet.net; Robert Clarke - clarker@georgetown.edu; Viktor Nazarov - vn8@georgetown.edu

* Corresponding author

Published: 02 May 2006

*BMC Cancer*2006, **6**:111 doi:10.1186/1471-2407-6-111

Received: 09 December 2005

Accepted: 02 May 2006

This article is available from: http://www.biomedcentral.com/1471-2407/6/111

Abstract

Background: The poly Q polymorphism in *AIB1* (amplified in breast cancer) gene is usually assessed by fragment length analysis which does not reveal the actual sequence variation. The purpose of this study is to investigate the sequence variation of poly Q encoding region in breast cancer cell lines at single molecule level, and to determine if the sequence variation is related to *AIB1* gene amplification.

Methods: The polymorphic poly Q encoding region of *AIB1* gene was investigated at the single molecule level by PCR cloning/sequencing. The amplification of *AIB1* gene in various breast cancer cell lines were studied by real-time quantitative PCR.

Results: Significant amplifications (5–23 folds) of *AIB1* gene were found in 2 out of 9 (22%) ER positive cell lines (in BT-474 and MCF-7 but not in BT-20, ZR-75-1, T47D, BT483, MDA-MB-361, MDA-MB-468 and MDA-MB-330). The *AIB1* gene was not amplified in any of the ER negative cell lines. Different passages of MCF-7 cell lines and their derivatives maintained the feature of *AIB1* amplification. When the cells were selected for hormone independence (LCC1) and resistance to 4-hydroxy tamoxifen (4-OH TAM) (LCC2 and R27), ICI 182,780 (LCC9) or 4-OH TAM, KEO and LY 117018 (LY-2), *AIB1* copy number decreased but still remained highly amplified. Sequencing analysis of poly Q encoding region of *AIB1* gene did not reveal specific patterns that could be correlated with *AIB1* gene amplification. However, about 72% of the breast cancer cell lines had at least one under represented (<20%) extra poly Q encoding sequence patterns that were derived from the original allele, presumably due to somatic instability. Although all MCF-7 cells and their variants had the same predominant poly Q encoding sequence pattern of $(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$ of the original cell line, a number of altered poly Q encoding sequences were found in the derivatives of MCF-7 cell lines.

Conclusion: These data suggest that poly Q encoding region of *AIB1* gene is somatic unstable in breast cancer cell lines. The instability and the sequence characteristics, however, do not appear to be associated with the level of the gene amplification.

Background

While predisposition to breast cancer is largely due to mutations in high penetrance tumor suppressor genes such as *BRCA1* and *BRCA*2, progression of cancer is the result of accumulation of genetic alterations. These alterations include gene amplifications, microsatellite instabilities, loss of heterozygosity, and mutations in genes that play important roles in signal transduction or transcription activation pathways leading to tumorigenesis. Gene amplification in breast cancer was found in several chromosomal locations [1-4]. Among them, ErbB2 (or HER-2/neu) amplification strongly correlates with steroid receptor negative tumors [5,6], and amplification of *AIB1*(amplified in breast cancer 1) gene is prevalent in estrogen receptor (ER) positive tumors [7,8]. The *AIB1* gene is a member of the SRC-1 (steroid receptor coactivator) family and is also known as RAC3, TRAM-1 or ACTR [7,9,10]. It is located at chromosome 20q12 region and encodes a protein of 1420 amino acids containing bHLH-PAS dimerization domain, a hormone receptor interaction domain, a CBP interaction domain, and histone acetyltransferase domain [11]. The amplifications and overexpression of *AIB1* gene were found to be a common phenomenon in breast cancer cell lines and primary breast cancer tissues [12-15]. Since *AIB1* bridges between nuclear receptors and other coactivators or the transcriptional machinery, its amplification and overexpression may play crucial roles in the development of breast cancer and may potentially have influence on the hormonal prevention and treatment for breast cancer.

Toward the C-terminus of AIB1, there is a stretch of polyglutamine residues that are encoded by polymorphic CAG repeats. The expansion of CAG repeats in poly Q containing proteins underlies a number of neurodegenerative diseases [16,17]. Large expansion of triplet repeats in *AIB1* gene does not occur, presumably due to the frequent interruption by CAA [18]. However somatic instability by nucleotide substitution such as small insertion or deletion does occur [18]. In androgen receptor (AR), the length of the CAG repeats inversely correlates with its transcriptional activity [19,20]. Meanwhile a shorter CAG repeat in AR is associated with a higher risk of an aggressive prostate cancer phenotype characterized by extraprostatic extension, distant metastases, or poor histological grade [21]. In the case of AIB1, it is not clear if the polymorphic length of poly Q affects the transactivation activity of AIB1. AIB1 interacts with ER in a ligand-dependent manner [7]. It also interacts with non-steroid nuclear receptors and transcription co-integrators such as thyroid and retinoid receptors and CBP-dependent transcription complexes [22,23]. Thus, amplification of *AIB1* gene impacts on both estrogen dependent and estrogen independent mechanisms leading to tumorigenesis [24-26]. Although antiestrogens are the most common type of endocrine therapy in breast cancer treatment, acquired resistance can be a major problem in clinical management of initially responsive breast cancer patients.

Understanding of the quantitative and qualitative changes of *AIB1* gene in estrogen-independent and antiestrogen resistant breast cancer cell lines may help in the selection of steroid or non-steroid antiestrogen therapies. Evaluation of *AIB1* gene amplification in previous reports is performed by FISH or Southern blot analysis [2,4,7]. In this report, we use the real-time quantitative PCR (Q-PCR) technique to assess the amplification of *AIB1* gene in various breast cancer cell lines and primary breast tumors. We also analyze the sequence characteristics and instability of the polymorphic poly Q encoding region at the single molecule level by cloning and sequencing of the DNA region containing CAG repeats.

Methods

Samples and DNA preparation

Primary breast tumor specimens with matching normal breast tissue samples were obtained from Fu Jen Catholic University, and Cardinal Tien Hospital, Taiwan, after surgical removal of the tumor according to the IRB approved protocol. The ER positive breast cancer cell lines were obtained from Georgetown University Lombardi Comprehensive Cancer Center Tissue Culture shared resource and American Type Cell Culture. A total of 25 cancer and 4 non-cancer breast tissue cell lines were studied. MCF-7 variants include different passages, MCF-7 p19, MCF-7 p72 and MCF-7 derivatives: LCC1 (selected for growth *in vitro* without estrogens) [27], LCC2 (selected from LCC1 by treatment with non-steroid antiestrogen 4-OH TAM) [28], LCC9 (selected from LCC1 by treatment with steroid antiestrogen ICI 182,780) [29], LY-2 (resistant to 4-OH TAM, KEO and LY 117018) and R27 (resistant to 4-OH TAM). AK-47 is derived from parental ER positive cell line ZR-75-1 with the loss of expression of ER. LCC6 is a more aggressive form of MDA-MB-435 [30]. A1N4 is a normal breast cell line that is ER negative. DNA from tissues, blood and cell cultures was extracted by salting out method [31].

Preparation of standard DNA for quantitative PCR

A region of 439 bp from exon 5 of *AIB1* gene was amplified with the forward primer; 5'-CAAGCGATCAAATGAGGGTAG-3' and the reverse primer; 5'-CATTGTTTCATATCTCTGGCG-3'. A fragment of 85 bp from 3' untranslated region of β_2-*microglobulin* gene (β_2-*M*) was amplified with the forward primer; 5'-TGCTGTCTCCATGTTTGATGTATCT-3' and the reverse primer; 5'-TCTCTGCTCCCCACCTCTAAGT-3' [32-34]. These PCR products were cloned into the pCR 2.1-TOPO vector (Invitrogen). The plasmid DNA was isolated and quantified using the DU640 Spectrophotometer (Beckman, Fuller-

ton, CA, USA). The copy numbers were calculated from absorbance at 260 nm and based on the molecular weight of the resulting plasmid. The plasmid DNAs were serially diluted over four logs to establish the standard curve giving a range from 400,000 to 40 copies/μl. In additional set of experiments the standard curve was constructed using genomic DNA prepared as a pool of equal amounts of blood DNA from 7 control individuals with normal *AIB1* copy number. 'Normal' genomic DNA (100 ng/μl) was diluted in water over four logs. Since 1 ng of genomic DNA contains approximately 330 copies of a single copy gene, five standards used range from 33000 to 3.3 copies/μl.

Real-time quantitative PCR (RT Q-PCR)

In real-time Q-PCR analysis, the primers used were 5'-GAGTTTCCTGGACAAATGAG-3' (forward) and 5'-CATTGTTTCATATCTCTGGCG-3' (reverse) for *AIB1* gene (Exon 5), and the same primers as used for standard DNA preparation for β_2-*M* gene, yielding 134 bp and 85 bp PCR products, respectively. The TaqMan probes were FAM-5' GCCGTATGTTGATGAAAACACCACA 3'-TAMRA and VIC-5' TTGCTCCACAGGTAGCTCTAGGAGG 3'-TAMRA, for *AIB1* and β_2-*M* gene respectively, each labeled with FAM or VIC (reporter dye) at the 5' end and TAMRA (quencher dye) at the 3' end. Each 10 μl real time Q-PCR reaction mixture contained 1 × TaqMan Universal PCR Master Mix (Applied Biosystems, Foster City, CA), 10 ng of genomic DNA, 0.3 μM of each primer, and 0.1 μM probe. The actin gene was also used as a reference. However, since the actin gene has multiple homologous copies, the data presented here were referenced to β_2-*M* gene. The amplification was carried out according to the conditions suggested by the manufacturer (initial denaturation at 95°C for 10 min and 40 cycles of 95°C for 15 s and 60°C for 1 min) using an ABI Prism 7700 Sequence Detection System (Applied Biosystems, Foster City, CA). Each measurement was performed in triplicate and the threshold cycle numbers (C_T) were measured. The copy number was generated from the C_T value and standard curve according to previously described procedures [32-34].

Cloning and sequencing

The poly Q containing fragment was amplified by the forward primer F: 5' GTCTTATACCTGGTGTATTG 3' and the reverse primer R: 5' CTGGGGGAAGCAGTCACATTAG 3', yielding a PCR product of 314 bp. The high fidelity amplification was carried out in a 30 μl reaction mixture containing 10 ng of genomic DNA, 0.2 μM of each primer, 1 × HF 2 PCR buffer, dNTPs, and Advantage-HF 2 polymerase according to the manufacturer's recommendation (Clontech Laboratories, Palo Alto, CA). After 1 min of initial denaturation at 94°C, the DNA was amplified by 30 cycles of 45 s at 95°C, 45 s at 55°C and 45 s at 72°C, followed by a final extension at 72°C for 5 min. The PCR products were purified and cloned into pCR2.1-TOPO (Invitrogen) vector according to the manufacturer's protocol. At least 8 clones from each sample were picked for sequencing using BigDye sequencing kit and analyzed on an ABI 377 DNA Sequencer (Applied Biosystems, Foster City, CA). Two primers, F and F2 (5' AGCAGGGTTTTCTTAATGCTC 3') were used for sequencing and loading of reactions onto alternate lanes for easy tracking. The sequence results were analyzed using sequence analysis software version 3.4.

Results

Amplification of AIB1 gene

Real-time Q-PCR analysis allows the measurement of actual copy number of *AIB1* gene using a single copy β_2-*microglobulin* gene as a reference. From the threshold cycle number and the standard curve, the ratio of the copy number of *AIB1* gene to that of β_2-*M* gene can be calculated. This ratio can be used as a measure of the amplification of the *AIB1* gene. The average copy number ratio of the *AIB1*/β_2-*M* in the blood samples from 48 age matched control individuals is determined to be 1.16 ± 0.38. An *AIB1*/β_2-*M* ratio above 2 SD of the mean (1.16 + 2 × 0.38 = 1.92) is defined as truly amplified. In addition, all measurements were repeated using a pool of normal genomic DNAs for standard curve construction. The results obtained using both methods were practically identical.

We first evaluate the amplification of *AIB1* gene in 26 primary breast tumors (13 ER positive and 13 ER negative) and corresponding surrounding normal breast tissue samples. *AIB1* gene was found to be amplified in 1 ER positive tumor sample that constitutes 3.8% of total or 7.6% of ER positive tumors. This result is consistent with previous report [3,7,13].

As shown in Table 1, 9 out of 29 cell lines had elevated *AIB1* at 2SD above the mean. All of them were ER positive. *AIB1* gene was not amplified in 7 ER positive cell lines: BT20, ZR-75-1, T47D, BT483, MDA-MB-361, MDA-MB-468 and MDA-MB-330. None of the 13 ER negative cell lines showed significant *AIB1* amplification. Some cell lines are derivatives of others. For example, AK-47 is derived from ER positive ZR-75-1 cell line. In AK-47 cells, loss of ER expression did not have any effect on *AIB1* copy number. The amplification of *AIB1* gene in different passages of ER positive MCF-7 cell lines remains at high levels of 18–23 fold of control. Loss of estrogen dependence in LCC1 is accompanied by moderate decrease in *AIB1* gene amplification (13.5 fold in LCC1 versus 22.2 fold in MCF-7). The level of *AIB1* gene amplification was reduced to 14.6 fold when the estrogen-independent cells became resistant to antiestrogen 4-OH TAM treatment (LCC2). Similar decrease in amplification (15.6 fold of control) of *AIB1* gene was observed in estrogen-independent cell line

Table 1: The ER status, *AIB1* amplification, and poly Q encoding sequences in breast cell lines

Cell line [a]	ER status	*AIB1* copy number ratio (tumor/normal)	Number of s equence patterns
MCF-10A	-	1.1	3
MCF-10A neo	-	1.3	3
A1N4[b]	-	2.0	2
AK-47	-	1.4	3
MDA-MB435	-	1.3	4
LCC6	-	2.0	7
MDA-MB157	-	1.1	6
MDA-MB134V	-	1.4	2
MDA-MB231N	-	1.2	2
HBL100	-	1.2	1
ZR-75-30	-	1.7	1
HS 578T	-	1.9	2
HS 578BST	-	1.1	2
MCF7	+	22.2	1
MCF-7 P19	+	18.7	3
MCF-7 P72	+	22.8	3
LY-2	+	19.9	2
R27	+	19.4	2
LCC1	+	13.5	3
LCC2	+	14.6	2
LCC9	+	15.6	2
BT474	+	4.9	2
BT-483	+	1.9	4
BT20	+	1.7	2
MDA-MB468	+	1.7	2
ZR-75-1	+	1.5	2
MDA-MB361	+	1.3	2
T47D	+	1.0	5
MDA-MB330	+	0.9	2

[a]LCC1: estrogen independent and responsive which is selected for growth in vivo without estrogens; LCC2: selected from LCC1 by treatment with 4-OH TAM; LCC9: selected from LCC1 by treatment with ICI 182780, resistant to ICI182780 and 4-OH TAM:, LY-2: selected for resistance to 4-OH TAM, KEO and LY 117018; R27: able to grow in the presence of 4-OH TAM. LCC6 is the more aggressive variant of MD-MB435. AK-47 is the variant derived from ZR-75-1. A1N4 is a normal breast cell line.
[b]Cell lines listed in Table 2 are in bold.

that has gained antiestrogen resistance to both non-steroid, 4-OH TAM, and steroid, ICI 182,780 antiestrogens (LCC9) (Table 1).

Somatic instability of poly Q encoding region of AIB1 gene in breast cancer cell lines

The polymorphic poly Q encoding region of *AIB1* contains CAG repeat that is frequently interrupted by CAA's. The poly Q region is part of the histone acetyltransferase domain. It is also where the recruitment and interactions with other components of the transcription activator complex takes place. In order to investigate if qualitative alteration in this region accompanied the quantitative change of *AIB1* gene in breast cancer cell lines, we cloned and sequenced the poly Q encoding region of the gene. The cloning/sequencing technique resolved the heterogeneous poly Q encoding sequences into distinct sequences, thus allowing the analysis at the single molecule level. At least 8 clones from each cell line were selected and sequenced. Theoretically, there should be only 2 distinct sequence patterns if the cell line is heterozygous for *AIB1* allele and one distinct sequence pattern if it is homozygous. However, 18/25 (72%) (data partially shown in Table 2) of the cell lines contain at least one poly Q encoding sequence pattern that represents less than 20% of the sequenced clones of the cell line. These results suggest that the under-represented sequences probably arise from the parental sequence by somatic mutation. Indeed these rare sequences differ from their parental sequence by one base pair substitution (CAG to CAA) or by insertion or deletion of CAGs. The high degree (72% of the cell lines) of somatic instability is probably characteristic for cancer cell lines since it only occurs in less than 5% (2/43) of the normal controls. We analyzed normal A1N4 cell line at two different times. The first time, ten clones of A1N4 cell line were sequenced. Pattern 2 (Table 2), $(CAG)_6CAA(CAG)_9(CAACAG)_3(CAACAG\text{-}CAG)_2CAA$, was found in 4 clones, and pattern 17, $(CAG)_4CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$ was found in 6 clones. The second time, 31 clones were sequenced. Sixteen had pattern 2 and 15 had pattern 17. There was no occurrence of "extra" poly Q encoding sequence. These results suggest that the occurrence of rare sequences is not due to cloning/PCR artifact.

Association between poly Q length or its specific encoding sequence with *AIB1* amplification was not recognized (Table 2). Somatic instability occurred in all variants of MCF-7 cell line, although they all maintain the parental allele as the predominant coding sequences (pattern 1). Two new sequences arise by insertion of 2 and 3 CAGs in passage 19. Another two new sequences occur in passage 72 by deletion of 1 and 2 CAG repeats. Similar somatic mutations occur in cell lines LCC1, LCC2, LCC9, LY-2 and R27. These mutations seem to occur randomly and independently in each cell line. There is not any single sequence that occurs more frequently than others, except pattern 5, which occurs 3 times by losing 1 CAG directly from the parental sequence.

AK47 was derived from ZR-75-1 by losing its ER activity. During the establishment of the cell line, additional somatic mutations occurred in the polyglutamine region (patterns 11 and 16). The poly Q encoding sequence of *AIB1* gene seems to be quite unstable in MDA-MB435 cell line. It has 4 distinct poly Q encoding sequence patterns with pattern 9 as the predominant one. Its variant LCC6 had a total of 7 different sequence patterns. These data are consistent with the genomic instability that is characteristic for cancer cells. Although poly Q encoding sequence

patterns do not seem to directly link to *AIB1* gene amplification, it is possible that the alteration in poly Q length affects protein-protein interaction, thus, the transactivation activity of *AIB1*. While most alterations do not change poly Q length significantly, rare sequence pattern in LCC2 with much shorter (only 14 repeats) poly glutamine tract may affect the co-transactivating activity of *AIB1* gene.

Discussion

Genetic and clinical phenotypic heterogeneity is the prominent characteristic of breast cancer. Multiple genetic alterations contribute to breast cancer development and progression [35,36]. The occurrence of DNA amplifications in breast cancer had been studied by Southern blot [1], FISH (fluorescence in situ hybridization) [4] and CGH methods (comparative genomic hybridization) [37-39]. We developed real time quantitative PCR method to more accurately assess the amplification of *AIB1* gene in breast cancer cell lines. Amplification of *AIB1* in breast cancer cell lines; BT-474 and MCF-7 were first reported by Guan et al. [3]. By FISH analysis, Anzick et al. [7] observed >20 fold amplification of *AIB1* gene in three ER positive breast cancer cell lines (BT-474, MCF-7 and ZR-75-1) and, to a lesser extent, in 10% primary breast tumors. In this study we did not detect significant amplification in ZR-75-1 cell line. The discrepancy may be explained by a different source of the cell line or by spontaneous change of the cell line during passages. In addition, FISH analysis is

Table 2: Poly Q sequence patterns and *AIB1* amplification level in MCF7 and its variants

Cell line[a]	amplification	Sequence patterns	Pattern	(Q)n	Frequency
MCF-7	22.3	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	7/7
MCF-7 P19	18.7	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	7/9
		$(CAG)_6CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	2	29	1/9
		$(CAG)_5CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	3	28	1/9
MCF-7 P72	22.8	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	5/7
		$(CAG)_3CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	4	25	1/7
		$(CAG)_3CAA(CAG)_7(CAACAG)_3(CAACAGCAG)_2CAA$	5	24	1/7
LCC1	13.5	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	7/9
		$(CAG)_3CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	4	25	1/9
		$(CAG)_3CAA(CAG)_{14}(CAACAG)_2(CAACAGCAG)_2CAA$	6	29	1/9
LCC2	14.6	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	4/5
		$CAG(CAACAG)_3(CAACAGCAG)_2CAA$	7	14	1/5
LCC9	15.6	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	7/8
		$(CAG)_3CAA(CAG)_2CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	8	29	1/8
LY-2	19.9	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	9/10
		$(CAG)_6CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	9	28	1/10
R27	19.4	**$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$**	1	26	8/9
		$(CAG)_3CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	4	25	1/9
MDA-MB435	1.3	$(CAG)_3CAA(CAG)_2CAA(CAG)_{11}(CAACAG)_2(CAACAGCAG)_2CAA$	10	29	2/8
		$(CAG)_6CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	9	28	3/8
		$(CAG)_3CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	1	26	2/8
		$(CAG)_6CAA(CAG)_8(CAACAG)_2(CAACAGCAG)_2CAA$	11	26	1/8
LCC6	2.0	$(CAG)_3CAA(CAG)_2CAA(CAG)_{10}(CAACAG)_2(CAACAGCAG)_2CAA$	12	28	1/9
		$(CAG)_3CAA(CAG)_2CAA(CAG)_6(CAACAG)_3(CAACAGCAG)_2CAA$	13	26	1/9
		$(CAG)_6CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	9	28	3/9
		$(CAG)_3CAA(CAG)_2CAA(CAG)_{10}(CAACAG)_3(CAACAGCAG)_2CAA$	14	30	1/9
		$(CAG)_6CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	15	28	1/9
		$(CAG)_3CAA(CAG)_2CAA(CAG)_{11}(CAACAG)_2(CAACAGCAG)_2CAA$	10	29	1/9
		$(CAG)_3CAA(CAG)_2CAACAGCAA(CAG)_8(CAACAG)_2(CAACAGCAG)_2CAA$	16	28	1/9
ZR-75-1	1.5	$(CAG)_6CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	2	29	6/7
		$(CAG)_5CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	3	28	1/7
AK-47	1.4	$(CAG)_6CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	2	29	2/8
		$(CAG)_6CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	11	28	5/8
		$(CAG)_5CAA(CAG)_8(CAACAG)_3(CAACAGCAG)_2CAA$	15	27	1/8
A1N4	2.0	$(CAG)_6CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	2	29	20/41
		$(CAG)_4CAA(CAG)_9(CAACAG)_3(CAACAGCAG)_2CAA$	17	27	21/41

[a]LCC1: estrogen independent and responsive which is selected for growth in vivo without estrogens; LCC2: selected from LCC1 by treatment with 4-OH TAM; LCC9: selected from LCC1 by treatment with ICI 182780, resistant to ICI182780 and 4-OH TAM:, LY-2: selected for resistance to 4-OH TAM, KEO and LY 117018; R27: able to grow in the presence of 4-OH TAM. LCC6 is the more aggressive variant of MD-MB435. AK-47 is the variant derived from ZR-75-1. A1N4 is a normal breast cell line. The parental sequence patterns are in bold.

restricted to a few cells, whereas real time qPCR analysis measures the gene in the overall DNA extract. Glaeser et al. [2] used the quantitative differential PCR to determine the amplification level of *AIB1* and found no amplification in breast or endometrial carcinomas. These methods did not give actual copy numbers of the gene. In this study, we used real-time Q-PCR to determine the level of *AIB1* amplification. The ability of real-time Q-PCR to detect the fluorescent signal from degraded sequence specific TaqMan probe at the very beginning period of exponential stage offered an accurate way of DNA quantification. When compared to CGH, FISH and Southern blot analysis, this method has the advantage of high sensitivity, reproducibility, and efficiency.

If only the original cell lines were counted, the *AIB1* gene was amplified in 2 out of 9 ER positive and none of 10 ER negative cell lines. Higher degree (22%) of *AIB1* amplification in ER positive breast cancer cell lines may suggest the association between *AIB1*gene amplification and ER status. This is further supported by the observation that the *AIB1* amplification moderately decreased when cells became ER independent as LCC1, LCC2 and LCC9 consistent with the role of AIB1 in ER-dependent signaling. All MCF-7 variant cell lines maintain high level of *AIB1* gene amplification of the parental cells. Additional gain of resistance to an antiestrogen, ICI182,780, does not have significant effect on *AIB1* amplification (from LCC1 to LCC9). Similarly, resistance to 4-OH TAM is not consistently accompanied with the change in *AIB1* gene amplification. LY-2, R27, and LCC2 were all selected from estrogen dependent MCF-7 cells against high dose of 4-OH TAM. *AIB1* gene amplification in LY-2 and R27 cell lines remained almost unchanged, whereas in LCC2, *AIB1* gene amplification is moderately decreased. These data suggest that resistance to 4-OH TAM does not necessarily affect *AIB1* gene amplification. It should be noted that in LCC2 there is a short poly Q containing mutant *AIB1* and lower expression of the gene may be compensated by the increase in co-transactivation activity of the mutant protein. Our observation of variations in *AIB1* gene amplification in various derivatives of MCF-7 cell clines is consistent with the previous report in which several MCF-7 sublines were shown to have the capacity to generate clonal heterogeneity. This represents an important selective advantage in MCF-7 in leading to aggressive and metastatic forms of the disease [40].

Besides the quantitative regulation of *AIB1* gene in breast cancer cell lines, the *AIB1* gene contains CAG repeat region which is a target for genetic instability in tumor progression. Large expansion of triplet repeat occurs in various neurodegenerative diseases [41-43]. These abnormal proteins form large aggregates that have been shown to tie up transcription factors that bind poly Q such as CREB [44]. The poly Q tract in the androgen receptor (AR) gene is unique in that the large expansion of poly Q encoding CAG repeat causes the X-linked spinal bulbar muscular atrophy (SBMA, or Kennedy Disease) [19,20], but short poly Q of AR is correlated with hormone-dependent transactivation [19,20] and more aggressive form of cancer [21,45]. *AIB1* shares several structural/functional similarities with AR. Both genes are involved in nuclear receptor mediated regulation of gene expression. Due to frequent interruption with CAA's, large expansion of the triplet was not observed in *AIB1* gene. This region, however, was quite unstable as evidenced by frequent CAA/CAG changes and small insertions and deletions. The PCR/cloning strategy allows us to investigate the polymorphic poly Q encoding region at single molecule level. Since we only sequenced a small number of clones from each individual, we cannot exclude the possibility that the under-represented alleles may be lost in PCR/cloning/sequencing process. Several distinct poly Q encoding sequence patterns were observed in LCC-6, T47D, and MDA-MB157. T47D is a cell line of notable genetic instability that was observed in the estrogen receptor gene [46]. LCC6 cell line which formed ascites was a more aggressive form of MDA-MB435 [30]. Since various rare poly Q encoding sequences seem to arise randomly and independently in different cell lines regardless of the *AIB1* gene amplification levels, we attribute these somatic mutations to genetic mutability in cancer cells.

Conclusion

The poly Q encoding sequence of *AIB1* gene is genetically unstable and is an easy target for somatic mutations in cancer cells.*AIB1* gene amplification occurs in only a small fraction of ER positive primary breast tumors and breast cancer cell lines. AIB1 gene amplification has not been found in ER negative primary tumor or breast cancer cell lines. Gain of estrogen independence and resistance to steroid antiestrogen may be accompanied by moderate decrease of *AIB1* gene amplification.

Abbreviations

4-OH TAM = 4-hydroxy tamoxifen; ACTR = Activator of Retinoid and Thyroid Receptors; *AIB1 = Amplified in Breast Cancer* gene *1*; bHLH-PAS = basic helix-loop-helix -Per-ARNT-Sim; β_2*-M* = β_2*-microglobulin* gene; CBP = CREB binding protein; CREB = cAMP-responsive element binding protein; ER = estrogen receptor; FAM = 6-Carboxy-Fluorescein; RAC3 = Receptor-Associated Co-activator 3; TRAM-1 = Thyroid Hormone Receptor Activator Molecule1; TAMRA = 6-Carboxy-Tetramethyl rhodamine

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

LW participated in the conception and design of study, acquisition, analysis, and interpretation of results, and the final write up of the manuscript. PD carried out the cloning/sequencing experiment. JL extracted the DNA. ML collected tumors and specimens. RC provided various cell lines and participated in discussion. VN performed the real time quantitative PCR analysis. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Mr. Song-Ping Wang for technical assistance. This project was supported by DOD Breast Cancer Research Program DAMD17-01-1-0257.

References

1. Courjal F, Cuny M, Simony-Lafontaine J, Louason G, Speiser P, Zeillinger R, Rodriguez C, Theillet C: **Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups.** *Cancer Res* 1997, **57:**4360-4367.
2. Glaeser M, Floetotto T, Hanstein B, Beckmann MW, Niederacher D: **Gene amplification and expression of the steroid receptor coactivator SRC3 (AIB1) in sporadic breast and endometrial carcinomas.** *Horm Metab Res* 2001, **33:**121-126.
3. Guan XY, Xu J, Anzick SL, Zhang H, Trent JM, Meltzer PS: **Hybrid selection of transcribed sequences from microdissected DNA: isolation of genes within amplified region at 20q11–q13.2 in breast cancer.** *Cancer Res* 1996, **56:**3446-3450.
4. Tanner MM, Tirkkonen M, Kallioniemi A, Isola J, Kuukasjarvi T, Collins C, Kowbel D, Guan XY, Trent J, Gray JW, Meltzer P, Kallioniemi OP: **Independent amplification and frequent co-amplification of three nonsyntenic regions on the long arm of chromosome 20 in human breast cancer.** *Cancer Res* 1996, **56:**3441-3445.
5. Zeillinger R, Kury F, Czerwenka K, Kubista E, Sliutz G, Knogler W, Huber J, Zielinski C, Reiner G, Jakesz R: **HER-2 amplification, steroid receptors and epidermal growth factor receptor in primary breast cancer.** *Oncogene* 1989, **4:**109-114.
6. Courjal F, Louason G, Speiser P, Katsaros D, Zeillinger R, Theillet C: **Cyclin gene amplification and overexpression in breast and ovarian cancers: evidence for the selection of cyclin D1 in breast and cyclin E in ovarian tumors.** *Int J Cancer* 1996, **69:**247-253.
7. Anzick SL, Kononen J, Walker RL, Azorsa DO, Tanner MM, Guan XY, Sauter G, Kallioniemi OP, Trent JM, Meltzer PS: **AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer.** *Science* 1997, **277:**965-968.
8. Guan XY, Meltzer PS, Dalton WS, Trent JM: **Identification of cryptic sites of DNA sequence amplification in human breast cancer by chromosome microdissection.** *Nat Genet* 1994, **8:**155-161.
9. Chen H, Lin RJ, Schiltz RL, Chakravarti D, Nash A, Nagy L, Privalsky ML, Nakatani Y, Evans RM: **Nuclear receptor coactivator ACTR is a novel histone acetyltransferase and forms a multimeric activation complex with P/CAF and CBP/p300.** *Cell* 1997, **90:**569-580.
10. Li H, Gomes PJ, Chen JD: **RAC3, a steroid/nuclear receptor-associated coactivator that is related to SRC-1 and TIF2.** *Proc Natl Acad Sci U S A* 1997, **94:**8479-8484.
11. Leo C, Chen JD: **The SRC family of nuclear receptor coactivators.** *Gene* 2000, **245:**1-11.
12. Bouras T, Southey MC, Venter DJ: **Overexpression of the steroid receptor coactivator AIB1 in breast cancer correlates with the absence of estrogen and progesterone receptors and positivity for p53 and HER2/neu.** *Cancer Res* 2001, **61:**903-907.
13. Bautista S, Valles H, Walker RL, Anzick S, Zeillinger R, Meltzer P, Theillet C: **In breast cancer, amplification of the steroid receptor coactivator gene AIB1 is correlated with estrogen and progesterone receptor positivity.** *Clin Cancer Res* 1998, **4:**2925-2929.
14. Reiter R, Wellstein A, Riegel AT: **An isoform of the coactivator AIB1 that increases hormone and growth factor sensitivity is overexpressed in breast cancer.** *J Biol Chem* 2001, **276:**39736-39741.
15. Shibata A, Hayashi Y, Imai T, Funahashi H, Nakao A, Seo H: **Somatic gene alteration of AIB1 gene in patients with breast cancer.** *Endocr J* 2001, **48:**199-204.
16. Paulson HL: **Protein fate in neurodegenerative proteinopathies: polyglutamine diseases join the (mis)fold.** *Am J Hum Genet* 1999, **64:**339-345.
17. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH: **Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy.** *Nature* 1991, **352:**77-79.
18. Dai P, Wong LJ: **Somatic instability of the DNA sequences encoding the polymorphic polyglutamine tract of the AIB1 gene.** *J Med Genet* 2003, **40:**885-890.
19. Chamberlain NL, Driver ED, Miesfeld RL: **The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function.** *Nucleic Acids Res* 1994, **22:**3181-3186.
20. Kazemi-Esfarjani P, Trifiro MA, Pinsky L: **Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: possible pathogenetic relevance for the (CAG)n-expanded neuronopathies.** *Hum Mol Genet* 1995, **4:**523-527.
21. Giovannucci E, Stampfer MJ, Krithivas K, Brown M, Dahl D, Brufsky A, Talcott J, Hennekens CH, Kantoff PW: **The CAG repeat within the androgen receptor gene and its relationship to prostate cancer.** *Proc Natl Acad Sci U S A* 1997, **94:**3320-3323.
22. Takeshita A, Cardona GR, Koibuchi N, Suen CS, Chin WW: **TRAM-1, A novel 160-kDa thyroid hormone receptor activator molecule, exhibitsdistinct properties from steroid receptor coactivator-1.** *J Biol Chem* 1997, **272:**27629-27634.
23. Torchia J, Rose DW, Inostroza J, Kamei Y, Westin S, Glass CK, Rosenfeld MG: **The transcriptional co-activator p/CIP binds CBP and mediates nuclear-receptor function.** *Nature* 1997, **387:**677-684.
24. Lam HY: **Tamoxifen is a calmodulin antagonist in the activation of cAMP phosphodiesterase.** *Biochem Biophys Res Commun* 1984, **118:**27-32.
25. Nardulli AM, Greene GL, O'Malley BW, Katzenellenbogen BS: **Regulation of progesterone receptor messenger ribonucleic acid and protein levels in MCF-7 cells by estradiol: analysis of estrogen'seffect on progesterone receptor synthesis and degradation.** *Endocrinology* 1988, **122:**935-944.
26. Clarke R, van den Berg HW, Murphy RF: **Reduction of the membrane fluidity of human breast cancer cells by tamoxifen and 17 beta-estradiol.** *J Natl Cancer Inst* 1990, **82:**1702-1705.
27. Brunner N, Boulay V, Fojo A, Freter CE, Lippman ME, Clarke R: **Acquisition of hormone-independent growth in MCF-7 cells is accompanied by increased expression of estrogen-regulated genes but without detectable DNA amplifications.** *Cancer Res* 1993, **53:**283-290.
28. Brunner N, Frandsen TL, Holst-Hansen C, Bei M, Thompson EW, Wakeling AE, Lippman ME, Clarke R: **MCF7/LCC2: a 4-hydroxytamoxifen resistant human breast cancer variant that retains sensitivity to the steroidal antiestrogen ICI 182,780.** *Cancer Res* 1993, **53:**3229-3232.
29. Brunner N, Boysen B, Jirus S, Skaar TC, Holst-Hansen C, Lippman J, Frandsen T, Spang-Thomsen M, Fuqua SA, Clarke R: **MCF7/LCC9: an antiestrogen-resistant MCF-7 variant in which acquired resistance to the steroidal antiestrogen ICI 182,780 confers an early cross-resistance to the nonsteroidal antiestrogen tamoxifen.** *Cancer Res* 1997, **57:**3486-3493.
30. Leonessa F, Green D, Licht T, Wright A, Wingate-Legette K, Lippman J, Gottesman MM, Clarke R: **MDA435/LCC6 and MDA435/LCC6MDR1: ascites models of human breast cancer.** *Br J Cancer* 1996, **73:**154-161.
31. Lahiri DK, Nurnberger JI Jr: **A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies.** *Nucleic Acids Res* 1991, **19:**5444.
32. Bai RK, Wong LJ: **Detection and quantification of heteroplasmic mutant mitochondrial DNA by real-time amplification refractory mutation system quantitative PCR analysis: a single-step approach.** *Clin Chem* 2004, **50:**996-1001.

33. Bai RK, Wong LJ: **Simultaneous detection and quantification of mitochondrial DNA deletion(s), depletion, and over-replication in patients with mitochondrial disease.** *J Mol Diagn* 2005, **7:**613-622.
34. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biol* 2002, **3:**RESEARCH0034.
35. Devilee P, Cornelisse CJ: **Somatic genetic changes in human breast cancer.** *Biochim Biophys Acta* 1994, **1198:**113-130.
36. Bieche I, Champeme MH, Lidereau R: **Loss and gain of distinct regions of chromosome 1 q in primary breast cancer.** *Clin Cancer Res* 1995, **1:**123-127.
37. Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldman FM: **Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization.** *Proc Natl Acad Sci U S A* 1994, **91:**2156-2160.
38. Muleris M, Almeida A, Gerbault-Seureau M, Malfoy B, Dutrillaux B: **Identification of amplified DNA sequences in breast cancer and their organization within homogeneously staining regions.** *Genes Chromosomes Cancer* 1995, **14:**155-163.
39. Muleris M, Almeida A, Gerbault-Seureau M, Malfoy B, Dutrillaux B: **Detection of DNA amplification in 17 primary breast carcinomas with homogeneously staining regions by a modified comparative genomic hybridization technique.** *Genes Chromosomes Cancer* 1994, **10:**160-170.
40. Nugoli M, Chuchana P, Vendrell J, Orsetti B, Ursule L, Nguyen C, Birnbaum D, Douzery EJ, Cohen P, Theillet C: **Genetic variability in MCF-7 sublines: evidence of rapid genomic and RNA expression profile modifications.** *BMC Cancer* 2003, **3:**13.
41. La Spada AR, Paulson HL, Fischbeck KH: **Trinucleotide repeat expansion in neurological disease.** *Ann Neurol* 1994, **36:**814-822.
42. Sinden RR: **Biological implications of the DNA structures associated with disease-causing triplet repeats.** *Am J Hum Genet* 1999, **64:**346-353.
43. Warren ST: **The expanding world of trinucleotide repeats.** *Science* 1996, **271:**1374-1375.
44. McCampbell A, Taylor JP, Taye AA, Robitschek J, Li M, Walcott J, Merry D, Chai Y, Paulson H, Sobue G, Fischbeck KH: **CREB-binding protein sequestration by expanded polyglutamine.** *Hum Mol Genet* 2000, **9:**2197-2202.
45. Kantoff P, Giovannucci E, Brown M: **The androgen receptor CAG repeat polymorphism and its relationship to prostate cancer.** *Biochim Biophys Acta* 1998, **1378:**C1-5.
46. Graham ML 2nd, Krett NL, Miller LA, Leslie KK, Gordon DF, Wood WM, Wei LL, Horwitz KB: **T47DCO cells, genetically unstable and containing estrogen receptor mutations, are a model for the progression of breast cancers to hormone resistance.** *Cancer Res* 1990, **50:**6208-6217.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2407/6/111/prepub

Endocrine therapy resistance can be associated with high estrogen receptor α (ERα) expression and reduced ERα phosphorylation in breast cancer models

Barbara Kuske[1], Catherine Naughton[1], Kate Moore[1], Kenneth G MacLeod[1], William R Miller[1], Robert Clarke[2], Simon P Langdon[1] and David A Cameron[1]

[1]CRUK Cancer Research Centre and Academic Breast Unit, University of Edinburgh, Crewe Road South, Edinburgh EH4 2XR, UK, [2]Department of Oncology, Georgetown University, Washington, District of Columbia, USA

(Requests for offprints should be addressed to S P Langdon; Email: simon.langdon@cancer.org.uk)

B Kuske and C Naughton contributed equally to this work.

Abstract

Hormone-dependent estrogen receptor (ER)-positive breast cancer cells may adapt to low estrogen environments such as produced by aromatase inhibitors. In many instances, cells become insensitive to the effects of estrogen but may still retain dependence on ER. We have investigated the expression, function, and activation of ERα in two endocrine-resistant MCF-7 models to identify mechanisms that could contribute to resistance. While MCF-7/LCC1 cells are partially estrogen dependent, MCF-7/LCC9 cells are fully estrogen insensitive and fulvestrant and tamoxifen resistant. In both MCF-7/LCC1 and MCF-7/LCC9 cell lines, high expression of ERα was associated with enhanced binding to the trefoil factor 1 (TFF1) promoter in the absence of estrogen and increased transcription of TFF1 and progesterone receptor. In contrast to the observations derived from hypersensitive and supersensitive models, these cells were truly estrogen independent; nevertheless, removal of ERα by siRNA, or fulvestrant, a specific ER downregulator, inhibited growth indicating dependence on ERα. In the absence of estrogen, neither ERα Ser^{118} nor Ser^{167} were phosphorylated as frequently found in other ligand-independent cell line models. Addition of estrogen activated ERα Ser^{118} in MCF-7 and LCC1 cells but not in LCC9 cells. We suggest that the estrogen-independent growth within these cell lines is accounted for by high levels of ERα expression driving transcription and full estrogen independence explained by lack of ERα activation through Ser^{118}.

Endocrine-Related Cancer (2006) **13** 1121–1133

Introduction

Estrogen receptor α (ERα) is a major growth regulator for many breast cancers and has provided an exploitable target for therapy (Ali & Coombes 2002). Estrogen binding to ERα promotes conformational changes in the receptor leading to dimerization and attachment to DNA, generally at the site of conserved estrogen response elements in the promoter regions of target genes (Ali & Coombes 2002). Functional regulation of ERα is additionally mediated via phosphorylation of key residues in the activation function 1 (AF-1) domain of ERα including Ser^{118} and Ser^{167} and these influence both DNA binding and recruitment of cofactor molecules (reviewed in Lannigan 2003). The activation of ER involves crosstalk with other growth factor-signaling pathways. There is extensive evidence that activation of the mitogen-activated protein kinase (MAPK)-signaling cascade and the phosphoinositol 3 kinase (PI3-K) pathway phosphorylate ERα at Ser^{118} and Ser^{167}, via extracellular signal-regulated kinase (ERK)1/2 and Akt respectively (Bunone *et al.* 1996, Martin *et al.* 2000, Lannigan 2003). Transcriptional activation of ERα then involves a dynamic process where large transcription complexes incorporating co-activator proteins are assembled in an ordered and combinatorial manner

DOI:10.1677/erc.1.01257

Online version via http://www.endocrinology-journals.org

(Glass & Rosenfeld 2000, Metivier *et al.* 2003). Well-defined estrogen-regulated genes include trefoil factor 1 (TFF1)/pS2 (Masiakowski *et al.* 1982, Jakowlew *et al.* 1984) and progesterone receptor (PGR; Nardulli *et al.* 1988).

While tamoxifen has been the established form of treatment for ER-positive breast cancers for more than 20 years, other anti-estrogen strategies, notably aromatase inhibitors (Johnston & Dowsett 2003) and selective estrogen downregulators (SERDs), are increasingly being used (Robertson 2002). Despite initial responsiveness to these agents, most tumors eventually recur with acquired resistance (Clarke *et al.* 2001, 2003). Multiple mechanisms, dependent on the form of endocrine treatment, are involved in the development of resistance and, in many cases, these mechanisms remain unclear. During the acquisition of endocrine resistance, progressive changes are frequently observed, with ER-positive breast cancer cells progressing in a stepwise manner from a fully estrogen-sensitive phenotype to an estrogen-sensitive, but no longer dependent phenotype, to a fully resistant phenotype (Clarke *et al.* 2001, 2003).

With the increasing clinical use of aromatase inhibitors, such as letrozole, anastrazole, and exemestane which act by inhibiting estrogen synthesis (Johnston & Dowsett 2003), there has been great interest in how breast cancer cells can adapt to low estrogen environments and become resistant to the effects of these drugs. In most cases of acquired anti-estrogen resistance, expression of ERα is retained, suggesting that resistance involves either changed functionality or bypass of the receptor. Culturing breast cancer cells in estrogen-low conditions to produce long-term estrogen deprivation (LTED) has identified mechanisms of estrogen hypersensitivity and estrogen supersensitivity (Yue *et al.* 2002, Martin *et al.* 2003, 2005*a*,*b*, Santen *et al.* 2005). Estrogen hypersensitivity is characterized by the ability of cells to respond to levels of estrogen at concentrations 2–3 log lower than required to stimulate wild-type cells (Yue *et al.* 2002, Santen *et al.* 2005). This mechanism involves increased expression of ERα alongside enhanced phosphorylation of ERα Ser^{118} and is associated with activation of the ERK1/2 and PI3-K pathways. Estrogen supersensitivity, wherein cells are apparently estrogen independent, is a mechanism again associated with enhanced ERα expression, ERK activation, and activation of ERα Ser^{118} and involves ERα being supersensitized by growth factor activation (Martin *et al.* 2003, 2005*a*).

While higher levels of ERα expression are generally associated with enhanced estrogen response, in certain cases tumors expressing high levels of ERα can be insensitive to endocrine manipulation. High levels of ERα expression have been associated with increased proliferation rates (Black *et al.* 1983) and poor prognosis in breast cancer patients not receiving adjuvant therapy (Black *et al.* 1983, Thorpe *et al.* 1993). It has been suggested that a high level of ERα may lead to constitutive activation (Fowler *et al.* 2004). This mechanism has recently been demonstrated by Fowler *et al.* (2004, 2006) in a tetracycline-inducible ERα expression model of the MCF-7 cell line, wherein increased ERα expression resulted in aberrant promoter occupancy and gene activation in the absence of estrogen. The increased receptor activity required the amino-terminal domain and was not inhibited by tamoxifen, supporting the notion of AF-1 activation, yet was independent of $Ser^{104/106}$ and Ser^{118} phosphorylation (Fowler *et al.* 2004).

In these models, the expression of ERα is still critical to the response and it has been suggested that use of a SERD such as fulvestrant (faslodex, ICI 182 780) would be a beneficial strategy once resistance to aromatase inhibitors has developed (Johnston *et al.* 2005, Martin *et al.* 2005*b*). A number of laboratories are developing models of resistance to this agent to identify strategies that might be tried at the onset of resistance (Dowsett *et al.* 2005, Howell 2005, Johnston *et al.* 2005, Martin *et al.* 2005*b*, Nicholson *et al.* 2005, Normanno *et al.* 2005).

We have investigated two MCF-7 cell lines (MCF-7/LCC1 and MCF-7/LCC9), which have acquired estrogen insensitivity and with variable sensitivity to tamoxifen and fulvestrant to identify novel mechanisms of endocrine resistance that might arise in clinical specimens. The wild-type ER-positive MCF-7 breast cancer cell line is both estrogen dependent and responsive to anti-estrogens, such as tamoxifen and fulvestrant. The MCF-7/LCC1 (LCC1) cell line was derived from an MCF-7 xenograft, which had grown in a low estrogen environment in an immuno-deprived mouse and which was known to be estrogen independent but with a degree of estrogen sensitivity (Brunner *et al.* 1993). Treatment of the cell line with fulvestrant produced the MCF-7/LCC9 (LCC9) cell line which is fully resistant to both estrogen and fulvestrant (Brunner *et al.* 1997). A number of novel features of these lines were identified within this study and are reported here.

Materials and methods

Cell proliferation

MCF-7 cells were routinely grown in phenol red containing Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal calf serum (FCS), penicillin (100 units/ml), and streptomycin (100 (g/ml). LCC1 and LCC9 cells (source: Dr Robert Clarke, V T Lombardi Cancer Research Center, Georgetown University Medical School, Washington, DC, USA) were routinely kept in phenol-free containing DMEM supplemented with 5% dextran-activated charcoal-stripped fetal calf serum (DCC), penicillin (100 units/ml), streptomycin (100 (g/ml), and 2 mM glutamine. All cells were grown at 37 °C in 5% CO_2. To determine the effects of 17β-estradiol (E_2) and tamoxifen on cell proliferation, MCF-7 cells were seeded in six-well plates in phenol red containing DMEM with 10% fetal bovine serum (FBS) for 24 h. The media were changed to phenol red-free DMEM with 5% DCC for 48 h. The cells were then supplemented with media containing either 1 nM E_2, 1 μM tamoxifen or both. LCC1 and LCC9 cells were seeded in six-well plates in phenol red-free containing DMEM with 5% DCC and after 24 h supplemented with E_2 and/or tamoxifen. Cell growth was evaluated using a Coulter counter. Fulvestrant was a kind gift from Dr Alan Wakeling (AstraZeneca, Macclesfield, Cheshire, UK). For studies exploring growth in DMEM without serum, the sulforhodamine-B (SRB) colorimetric assay was used.

Briefly, log phase cells were seeded into 96-well flatbottom tissue culture plates. The following day, cells were washed in PBS and media replaced with phenol red-free DMEM for 48 h. Cells were then treated with concentrations of E_2 varying from 10 fM to 1 μM in the absence or presence of 100 nM fulvestrant. After 72 h, plates were removed from the incubator and ice-cold 25% trichloroacetic acid (TCA) solution (50 μl) added to each well. All plates were placed on ice for 60 min after which the TCA solution was removed. The plates were washed under running water and dried prior to staining with SRB dye solution (30 min at room temperature) and the trays were washed with 1% glacial acetic acid (×4) at room temperature, air-dried, and resuspended in 10 mM Tris buffer (pH 10.5; 150 μl) before reading at 540 nm.

RNA extraction and RT-PCR

Extraction of total RNA from whole cells was performed using Tri-Reagent (Sigma) as per the manufacturers' instructions. RNA concentration was measured using a spectrophotometer. QuantiTect SYBR Green system (Qiagen, cat no. 204243) was used according to the manufacturers instructions for one step RT-PCR in a total of 15 μl reaction volumes, including 0.5 μM each primer and 40 ng RNA. Real-time cycler conditions were RT: 50 °C for 30 min; PCR: initial activation 95 °C for 15 min followed by 40 cycles of denaturation 94 °C for 15 s, annealing 57 °C for 30 s, extension 72 °C for 30 s, and a final extension of 72 °C for 60 s. The following primers were used:

TFF1: fwd TTGTGGTTTTCCTGGTGTCA
rev CCGAGCTCTGGGACTAATCA
ERα: fwd CCACCAACCAGTGCACCATT
rev GTCTTTCCGTATCCCACCTTTC
PGR: fwd GTCAGTGGGCAGATGCTGTA
rev AGCCCTTCCAAAGGAATT
ACTIN: fwd CTACGTCGCCCTGGACTTCGAGC
rev GATGGAGCCGCCGATCCACACGG

Western analysis

Cells were washed twice with PBS and lysed in ice-cold lysis buffer (50 mM Tris (pH 7.5), 5 mM EDTA (pH 8.5), 150 mM NaCl, 1% Triton X-100, aprotinin 10 μg/ml, and 1× protease cocktail inhibitor (Roche) for 10 min and the debris was cleared by centrifugation at 13 000 r.p.m. for 6 min at 4 °C). Protein lysates (100 μg) were resolved on 7.5–12% SDS-PAGE and electrophoretically transferred to Immobilon-P membranes. After transfer, membranes were blocked and probed with primary antibody overnight at 4 °C. Immunoreactive bands were detected using chemiluminescent reagents (ECL or SuperLuminol) and photographic paper (Hyperfilm, Amersham). The following antibodies were used: ERα (F-10; Santa Cruz Biotech, Santa Cruz, CA, USA sc-8002), PGR (ab-8; Neomarkers, Stratech Scientific Ltd, Newmarket, Suffolk, UK (MS-298)), P-ERK1/2 (1:1000, Cell Signaling, New England Biolabs, Hitchin, Herts, UK #9101), phospho-Ser^{118} ERα (1:500, Cell Signaling #2511), phospho-Ser^{167} ERα (1:500, Cell Signaling #2514), and actin (1:120 000, CP01, Calbiochem, La Jolla, CA, USA). Integrated optical density absorbance values were obtained by densitometric analysis using a gel scanner and analyzed by 'Labworks' gel analysis software (UVP Life Sciences, Cambridge, UK).

Chromatin immunoprecipitation assays (ChIP)

Cells were grown to 85–90% confluence in phenol red-free DMEM with 5% DCC for at least 48 h. Cells were

cross-linked with 1% formaldehyde (37 °C for 10 min) at 10-min interval over a 90-min time course. Unreacted formaldehyde was quenched by gentle agitation at room temperature for 10 min with 0.125 M glycine. Cells were then washed twice with ice-cold PBS, collected into PBS containing protease inhibitors (Roche), and centrifuged for 4 min at 2000 r.p.m. at 4 °C. The pellets were resuspended in lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris–HCl (pH 8.1), and 1× protease inhibitor cocktail), incubated on ice for 10 min, and sonicated (12×20 s at two amplitude microns, Soniprep 150, MSE) to fragment DNA to ~500 bp. Following centrifugation for 15 min at 13 000 r.p.m. and 4 °C, supernatants were collected and resuspended in dilution buffer (0.01% SDS, 1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris–HCl (pH 8.1), 167 mM NaCl, and 1× protease inhibitor cocktail). Chromatin were precleared with 1 μg anti-rabbit or anti-mouse IgG, 2 μg sheared salmon sperm DNA, and Protein-G-Agarose (50 μl of 50% slurry in dilution buffer) for 3 h at 4 °C. Immunoprecipitation using Protein-G-Agarose Beads (Roche) was performed overnight at 4 °C with anti-ERα HC-20 antibody (sc-543, Santa Cruz). Beads were washed sequentially for 5 min each at 4 °C with TSE I (20 mM Tris (pH 8.1), 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, and 0.1% SDS), TSE II (20 mM Tris (pH 8.1), 2 mM EDTA, 500 mM NaCl, 1% Triton X-100, and 0.1% SDS), and buffer III (10 mM Tris (pH 8.1), 0.25 M LiCl, 1 mM EDTA, 1% NP40, and 1% deoxycholate). Precipitates were then washed twice with TE buffer and the protein/DNA complexes were eluted twice with 0.1 M $NaHCO_3$ and 1% SDS. Heat treatment at 65 °C overnight reversed formaldehyde cross-links. DNA fragments were purified using QIAquick Spin Kit columns (Qiagen) and amplified using the QuantiTect SYBR Green system (Qiagen, cat no. 204242). TFF1 PCR conditions were: initial activation of 95 °C for 15 min followed by 45 cycles of 94 °C for 15 s, 55 °C for 30 s, 72 °C for 30 s, and a final extension of 72 °C for 5 min. TFF1 primer sequences: fwd GACGGAATGGGCTTCATGAGC and rev CTGAGACAATAATCTCCACTG. For the distal region, primers were: fwd GAGTTTGGCCTCC-CACATTA and rev CTTGCCTCTGCATTCTCTCC.

Short interfering (siRNA) transfections

MCF-7 cells were seeded at 0.5×10^6 cells per T75 flask in DMEM as mentioned previously. After 24 h, the media were changed to phenol red-free containing DMEM with 5% DCC for 48 h. LCC1 and LCC9 cells were seeded directly into phenol red-free containing DMEM with 5% DCC for 24 h prior to transfection. Cells were transfected with siRNA for 4 h using Oligofectamine reagent (Invitrogen) after which time 1 nM E_2 was added for a further 48 h prior to RNA and protein extraction. For the 7-day time course, the media were left unchanged after the initial changes. For siRNA growth assays, cells were seeded as for growth characterization as mentioned previously. siRNA transfections were carried out as described earlier but scaled down for 24-well plates. Following siRNA treatment for 4 h, cells were treated with 1 nM E_2 or 100 nM fulvestrant or a combination and cell counts on days 0, 3, and 6 were estimated using a Coulter counter. The following siRNA sequences were used: ER RNAi 1; ESR1 SMARTpool (four pooled sequences; Upstate Biotechnology, Lake Placed, NY, USA; M-003401; 100 nmol), ER RNAi 2; 5′-AAACAGGAGGAA-GAGCTGCCA (Ambion; 40 nmol), ER RNAi 3; 5′-AACCTCGGGCTGTGCTCTTTT (Ambion, Hunting-don, Cambridgeshire, UK; 40 nmol), and negative RNAi: Upstate (D-001206; 100 nmol).

Results

Increased ERα expression in resistant cell lines

To explore the possibility that high ERα expression leads to estrogen-independent growth in endocrine-resistant cells, the expression levels of ERα in resistant lines (LCC1 and LCC9) were compared with levels in wild-type MCF-7 cells. Both resistant lines expressed between four- and elevenfold more ERα mRNA than wild-type cells (Fig. 1A). ERα protein levels were clearly elevated in LCC1 cells relative to MCF-7 cells (sevenfold) and less markedly in LCC9 cells (Fig. 1B). E_2 decreased ERα protein in MCF-7 cells at 48 h and this has been explained by proteosomal degradation, a process speculated to limit the action of estrogen signaling (Nawaz *et al.* 1999; Fig. 1C). Similarly, both resistant lines demonstrated ERα turnover, suggesting that ERα is binding to E_2 in all cases. In contrast, tamoxifen treatment results in maintenance of the receptor expression levels in all three cell lines (Fig. 1C).

Addition of 1 nM 17 β-estradiol (E_2) to MCF-7 cells produced a marked stimulation of growth to cells cultured in estrogen-depleted (double charcoal-stripped FCS) medium (Fig. 2A). In the absence of E_2, MCF-7 cells are essentially static (Fig. 2A). In contrast, LCC1 cells grow rapidly in estrogen-depleted conditions and show an approximately twofold stimulation of growth on addition of E_2 (Fig. 2B). LCC9 cells showed a lack of response to E_2, again

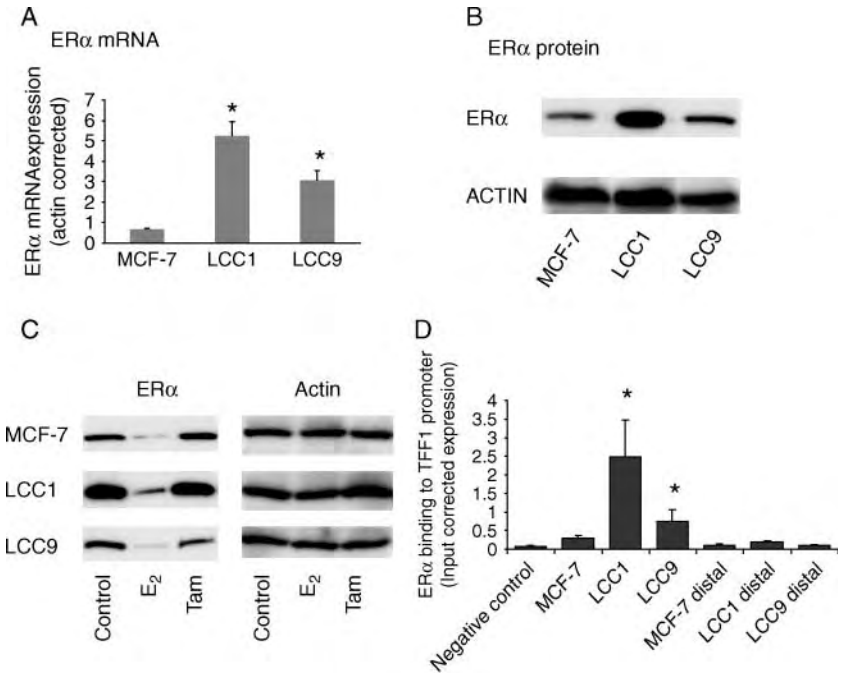


Figure 1 ERα expression in MCF-7, LCC1, and LCC9 cells. (A) ERα mRNA expression. Cells were grown in charcoal-stripped serum-containing medium for at least 48 h and RNA was collected. A representative experiment is shown of at least two experiments carried out. Each column presents mean of triplicate RT-PCR analysis for each sample demonstrating mRNA expression relative to actin expression. Error bars = S.D. Statistical significance noted for treatment groups versus matched control (one-way ANOVA and multiple comparison Tukey–Kramer test; $^*P<0.05$). (B) Western blot analysis of ERα (66 kDa) in breast cancer cell lines grown in charcoal-stripped serum-containing medium for 48 h prior to protein collection. One hundred micrograms of protein were loaded per lane and detected using either anti-ERα (Santa Cruz Biotech) or anti-actin (Calbiochem) antibodies as described in Materials and methods. (C) Western blot analysis of ERα (66 kDa) in breast cancer cell lines grown in charcoal-stripped serum-containing medium for at least 48 h prior to protein collection. One hundred micrograms of protein were loaded per lane and detected using either anti-ERα (Santa Cruz Biotech) or anti-actin (Calbiochem) antibodies as described in Materials and methods. (D) ERα binding to the TFF1 promoter. Basal recruitment of ERα to the TFF1 promoter was determined by ChIP analysis on untreated cells. The ChIP method used was as described in Materials and methods and immunoprecipitated TFF1 promoter was quantified by real-time PCR. The input-corrected expression values were determined by normalizing to the inputs. Data are presented as mean + S.E. Groups were compared with the Kruskal–Wallis test with Dunn's multiple comparison test ($^*P<0.05$). Binding to the promoter region is compared with binding to a region 3.5 kb distal to the promoter wherein only background binding was observed.

growing very rapidly in the absence of E_2 (Fig. 2C). Addition of 1 μM tamoxifen to MCF-7 cells antagonized the E_2-stimulated growth in this cell line. Tamoxifen also inhibited the E_2-stimulated growth of LCC1 cells but had no effect on LCC9 cells (Fig. 2B and C). These results are consistent with wild-type cells being estrogen dependent, LCC1 cells demonstrating partial estrogen dependence and LCC9 cells being fully estrogen independent.

Reduced ERα Ser^{118} phosphorylation in LCC9 cells

Several frequently cited mechanisms of estrogen-independent activation of ERα involve phosphorylation of ERα at the Ser^{118} or Ser^{167} residues mediated via ERK or Akt respectively (Bunone *et al.* 1996, Martin *et al.* 2000, Lannigan 2003). While the Ser^{118} residue is a major site of E_2-induced phosphorylation, Ser^{167} is not (Lannigan 2003). The latter site is activated by growth factor signaling. In view of these previous observations, we first investigated whether ERα Ser^{118} or Ser^{167} phosphorylation were increased in the absence of estrogen in the resistant cell lines. Neither was there evidence of increased Ser^{118} phosphorylation in the resistant lines relative to MCF-7 under basal conditions, nor was Ser^{167} phosphorylation increased (Fig. 3A–C). Furthermore, phospho-ERK1/2 expression was unchanged in the lines (Fig. 3C). On E_2 addition, there was a marked increase in Ser^{118} phosphorylation in MCF-7 cells and this was also observed in the LCC1 cell line (Fig. 3A and B). However, minimal change was observed on E_2 addition to LCC9 cells (Fig. 3A and B). Ser^{118} phosphorylation has been proposed to affect cofactor recruitment and this might explain the reduced

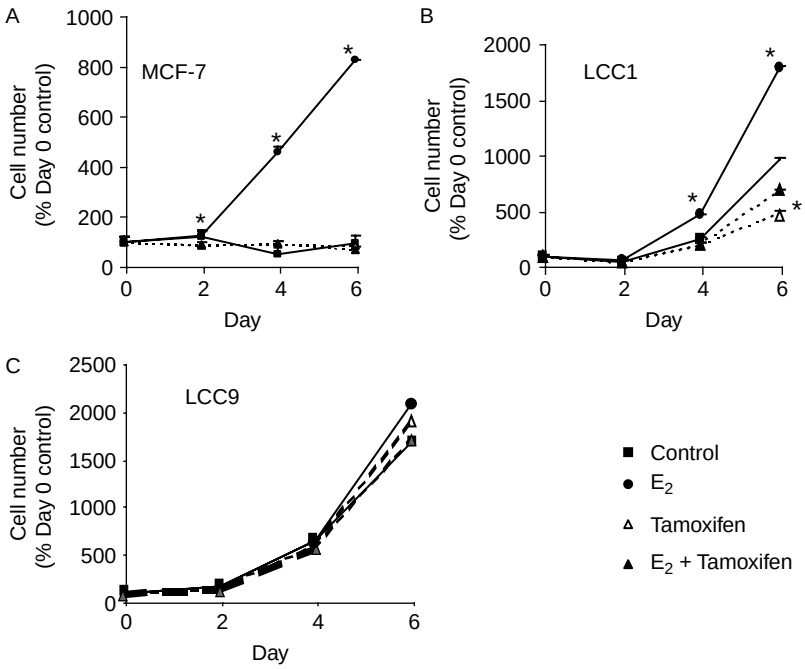


Figure 2 Growth characterization of MCF-7 and MCF-7 variant cells. (A) MCF-7 cells, (B) LCC1, and (C) LCC9 cells were plated for 24 h and maintained in reduced media for 48 h before treatment. Cells were then left untreated (control group), treated with 1 nM E_2, 1 μM tamoxifen or 1 nM E_2 and 1 μM tamoxifen. Cells were counted on day 0 (72 h after plating/day of treatment start) and days 2/4/6 using a Coulter counter. Mean cell counts of triplicate samples and duplicate counts for each time point in each treatment group are expressed. Error bars = S.D. A representative experiment is shown of at least four experiments carried out.

transcriptional (as mentioned below) and growth responses observed on E_2 addition to this cell line. Tamoxifen alone produced a small increase in Ser^{118} phosphorylation in MCF-7 and LCC1 cells but not in LCC9 cells (Fig. 3A and B). Tamoxifen also produced a reduction of estrogen's Ser^{118} phosphorylation in the MCF-7 and LCC1 cell lines (Fig. 3A and B).

Modified DNA binding of ERα in resistant cell lines

To explore whether high ERα expression was reflected in enhanced DNA binding in the absence of E_2, ChIP methodology was used to examine ERα binding to the promoter of the E_2-responsive gene TFF1 in the MCF-7, LCC1, and LCC9 cell lines. LCC9 cells had >2.5-fold greater ERα binding to the TFF1 promoter than MCF-7 cells (Fig. 1D). However, this binding was significantly higher in LCC1 cells with levels greater than eightfold above MCF-7 cells. This enhanced ERα binding in LCC1 cells was equivalent to the increased expression of ERα protein and is consistent with the suggestion by Fowler *et al.* (2004) that enhanced ERα protein expression can lead to increased DNA binding. As a control, binding to a region 3.5 kb distal to this region indicated only background levels as expected (Fig. 1D).

Growth responses to estrogen and tamoxifen in the wild-type and variant cell lines are reflected in transcriptional changes

To investigate the differences in estrogen and anti-estrogen activation processes, indicator genes that reflected the different growth responses were next investigated. Transcriptional changes in the estrogen-regulated genes TFF1 and PGR were measured and modulated expression was compared with the growth changes.

Expression of TFF1 mRNA in the absence of E_2 was higher in both resistant lines compared with MCF-7 cells (Fig. 4A). After 48-h E_2 (1 nM) treatment, TFF1 mRNA was increased by >20-fold in MCF-7 cells, but only one- to twofold in the resistant lines although this increase was significant. Tamoxifen (1 μM) produced a small increase in TFF1 expression in MCF-7 and LCC1 cells but not in the LCC9 cell line (Fig. 4A). These levels broadly reflect the growth differences observed.

The expression of PGR mRNA in the absence of E_2 was greater in LCC1 and LCC9 lines compared with MCF-7 cells (Fig. 4B). As for TFF1, after 48-h E_2 treatment, PGR mRNA was increased by >20-fold in MCF-7 cells and 2–5-fold in LCC1 and LCC9 cell lines (Fig. 4B). Tamoxifen also increased the PGR mRNA

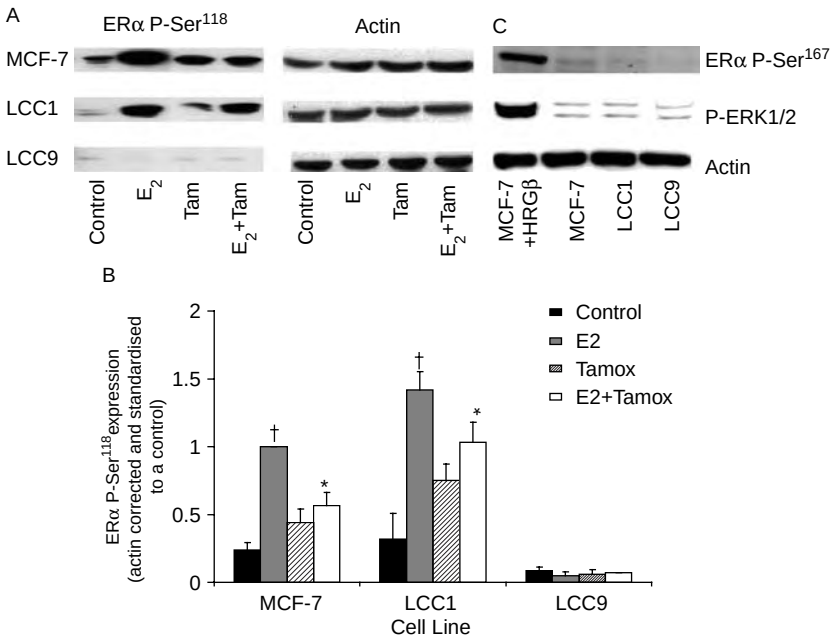


Figure 3 Effects of E_2 and tamoxifen on ERα phosphorylation in the resistant cell lines. (A) Western analysis of ERα phospho-Ser^{118} after 30-min treatment with control (no treatment), 1 nM E_2, 1 μM tamoxifen, or 1 nM E_2 and 1 μM tamoxifen. Lysates were run on a 10% SDS gel and membranes probed with anti-phospho ERα Ser^{118} antibody (1:1000). Lysates were also probed for actin expression to compare protein loading. (B) Histogram representing optical densities from triplicate western blots of ERα phospho-Ser^{118} after 30-min treatment with control (no treatment), 1 nM E_2, 1μM tamoxifen, or 1 nM E_2 and 1 μM tamoxifen. Values were actin corrected and then standardized to a control sample. The control sample was a 30 min E_2-treated MCF-7 sample and was used on all gels as a standard to allow comparisons between runs. Statistical comparisons were made with each cell line's control level; $^{*}P<0.01$; $^{\dagger}P<0.001$ (ANOVA). (C) Western analysis of ERα phospho-Ser^{167} and phospho-ERK1/2 in cell lines. Untreated lysates were probed with antibodies specific for ERα phospho-Ser^{167} and phospho-ERK1/2. A positive control lane of MCF-7 cells treated with 1 nM HRGβ was used. Lysates were also probed for actin expression.

expression level not only in MCF-7 cells, but also in LCC1 cells producing effects equivalent to that of E_2 in the latter cell line. No change was observed in the LCC9 cell line.

These results are consistent with transcription of TFF1 and PGR being increased by ligand-independent mechanisms in LCC1 and LCC9 cell lines with estrogen and tamoxifen producing an additional ligand-dependent increase.

Effect of removal of ERα on the growth of the cell lines

To determine the relative importance of ERα on downstream gene expression and growth of MCF-7, LCC1, and LCC9 cells, we investigated the effects of removing ERα, either by specific siRNA inhibition of receptor synthesis or through inhibition and degradation of the receptor by fulvestrant.

A panel of interfering RNAs (siRNAs) were initially compared for their ability to transiently reduce ERα expression and were transfected into the MCF-7 cell line. RNAi 1 is a pooled set of four targeted sequences (Imai *et al.* 2005) while RNAi 2 (5′-AAACAGGAG-GAAGAGCTGCCA) and RNAi 3 (5′-AACCT-CGGGCTGTGCTCTTTT) are individually targeted sequences (Leu *et al.* 2004). Of the three, RNAi 2 produced the best reduction of ERα mRNA and protein and was selected for further experiments (Fig. 5A and B). Quantitative RT-PCR analysis showed that, 48 h after transfection, ERα RNAi 2 treatment resulted in an 85% decrease in ERα mRNA expression and an 87% decrease in the presence of E_2 (Fig. 5C). LCC1 and LCC9 cells have significantly higher basal expression of ERα mRNA and siRNA removal caused an 82 and 73% decrease respectively with similar reductions in the presence of E_2 (Fig. 5C). Western analysis of the MCF-7 and LCC1 cell lines demonstrated that RNAi 2 produced ERα protein knockdown over a 7-day period (Fig. 5D) and it was effective in all three cell lines (Fig. 5E). This reduction in ERα protein was accompanied by a decrease in PGR protein (Fig. 5E). Thus, it appeared that gene expression in all three cell lines was ERα dependent.

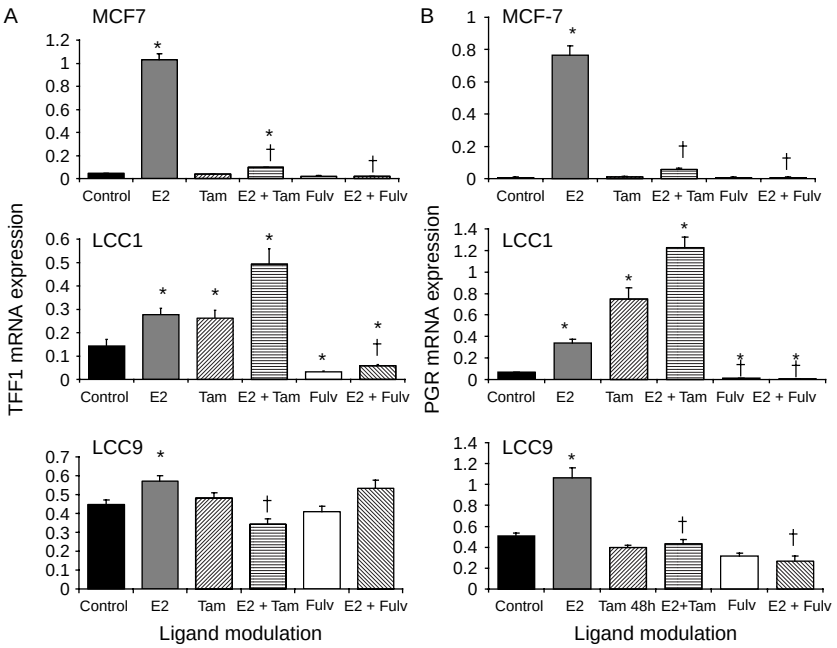


Figure 4 Effect of estrogen, tamoxifen, and fulvestrant on (A) TFF1 expression and (B) PGR mRNA expression in the cell lines. Relative mRNA expression values of TFF1 and PGR in the cell lines were measured by real-time RT-PCR using specific primer pairs. RNA was collected at 48 h and was extracted from either untreated (control) cells or cells treated with 1 nM E_2, 1 μM tamoxifen, 1 nM E_2+1 μM tamoxifen, 100 nM fulvestrant, or 1 nM E_2+100 nM fulvestrant. Each column represents mean of triplicate PCR analysis for each sample demonstrating mRNA expression relative to actin expression. Error bars=S.D. Statistical significance noted for treatment groups compared to matched control where $^*P<0.05$, untreated control versus treatment group; $^\dagger P<0.05$, E_2 control versus treatment group (ANOVA and multiple Tukey–Kramer comparison test).

This was investigated further using fulvestrant. Fulvestrant abrogates E_2-induced gene transcription by binding, blocking, and causing the degradation of ERα (Parker 1993). Fulvestrant treatment in MCF-7 cells blocked E_2-induced expression of TFF1 and PGR (Fig. 4A and B). In addition, ligand-independent and E_2-induced TFF1 and PGR expression in LCC1 cells were reduced on fulvestrant treatment. These data confirm that for LCC1 cells TFF1 and PGR induction are dependent on ERα expression. However, LCC9 cells are resistant to fulvestrant treatment and as such no change in TFF1 expression and only a minor change in PGR expression was observed. The effect of fulvestrant on the growth of all three cell lines was also investigated in the complete absence of serum (Fig. 6). Under these conditions, MCF-7 cells did not grow over a 72-h period. LCC1 cells, however, still proliferated and the addition of E_2 had little effect on growth confirming their independence of E_2. Under these conditions, fulvestrant was able to oppose the effect of low concentrations of E_2 again indicating dependence on ERα. In contrast, LCC9 cells were completely insensitive to both E_2 and fulvestrant. Fulvestrant degraded ERα protein in all three lines which is shown in Fig. 7A.

To determine how critical levels of ERα expression were for the growth of MCF-7, LCC1, and LCC9 cell lines, we used RNAi removal with or without fulvestrant to inhibit the synthesis of ERα protein (Fig. 7B–D). E_2-induced MCF-7 cell growth was significantly decreased (33%) by ERα removal and abolished by all combinations of fulvestrant alone or with RNAi. LCC1 cells grew in the absence of E_2 and RNAi removal had only a minor effect on growth. E_2-induced LCC1 cell growth was reduced by approximately 40% when ERα was removed through RNAi, but, unlike MCF-7 cells, fulvestrant alone was not enough to abolish growth – this, however, could be accomplished though through combination with RNAi. LCC9 cell growth in the absence of E_2 was reduced by ERα RNAi. A similar decrease was observed in the presence of E_2. LCC9 cells are fulvestrant resistant and no effect on growth was observed with this agent. No combinations of fulvestrant or RNAi were able to totally abolish growth. These results indicate a varying degree of dependence on ERα for growth in the three cell lines.

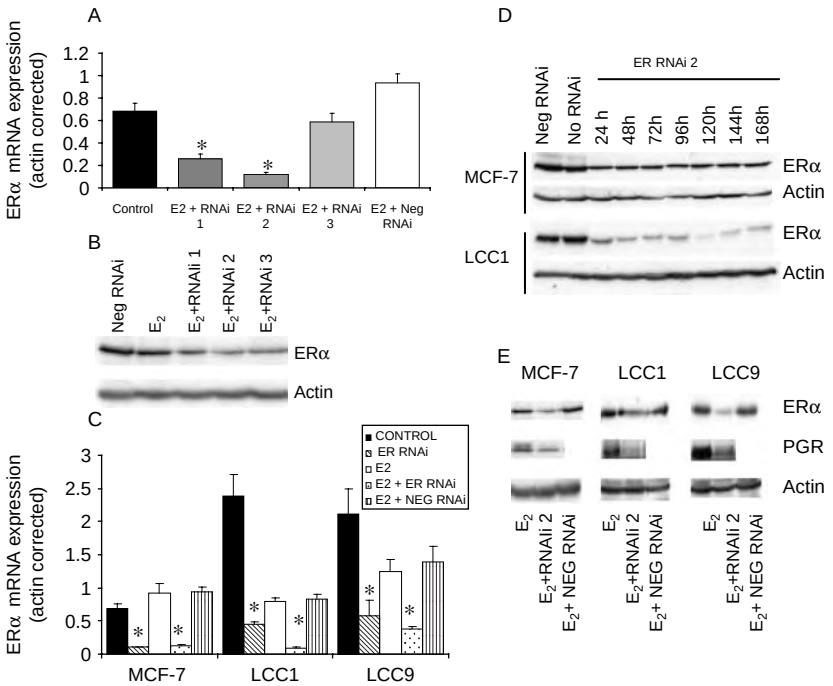


Figure 5 Effects of ERα RNAi on ERα and PGR expression in the cell lines. (A) Expression of ERα mRNA after treatment with a range of ERα mRNA-targeted RNAis. ERα mRNA expression was measured by quantitative RT-PCR of mRNA from MCF-7 cells 48 h after RNAi treatment in the presence of 1 nM E_2. Data are presented as mean ± S.D. of actin-corrected values from triplicate samples. The RNAi transfection method and RNAi sequences used are described in Materials and methods. Statistical significance noted for treatment groups compared with matched control where $^*P<0.05$, untreated control versus treatment group (ANOVA and multiple Tukey–Kramer comparison test). (B) Western analysis of ERα protein expression in MCF-7 cells 48 h after siRNA treatment. ERα was probed with the F-10 antibody and actin is shown as a loading control. (C) Expression of ERα mRNA after treatment with RNAi 2. ERα mRNA expression was measured by quantitative RT-PCR of mRNA from cell lines 48 h after RNAi 2 treatment in the presence or absence of 1 nM E_2. Data are presented as mean ± S.D. of actin-corrected values from triplicate samples. The RNAi transfection method and RNAi sequences used are described in Materials and Methods. Statistically significance differences are noted for treatment groups compared with matched control where $^*P<0.05$, untreated control versus treatment group (ANOVA and multiple Tukey–Kramer comparison test). (D) Western analysis time course of the effect of RNAi 2 treatment on ERα protein expression in the MCF-7 and LCC1 cell lines. ERα was probed with the F-10 antibody and actin is shown as a loading control. (E) Western analysis of ERα and PGR protein expression in the cell lines 48 h after siRNA treatment. ERα was probed with the F-10 antibody, PGR with Ab 8 and actin is shown as a loading control.

Discussion

Aromatase inhibitors are now used for the adjuvant treatment of most hormone receptor-positive early breast cancer. Despite the improvement they offer over tamoxifen alone, recurrences still occur, and thus models of resistance to both tamoxifen and estrogen deprivation are required. The series of MCF-7-derived cell lines provides an excellent model system for the exploration of mechanisms of stepwise acquisition of resistance to tamoxifen and estrogen deprivation. Most models to date have been derived *in vitro*, which makes LCC1 cells interesting as the initial estrogen deprivation was achieved *in vivo* and therefore might reflect features that could arise in a primary breast cancer (Brunner *et al.* 1993). In many of the *in vitro*-derived LTED models, acquired resistance is due to enhanced sensitization to low concentrations of estrogen, which often involves crosstalk with growth factor-signaling pathways (Martin *et al.* 2003, 2005*a*,*b*). LCC1 cells have certain of the characteristics of the LTED phenotype (Yue *et al.* 2002, Martin *et al.* 2003, 2005*a*,*b*, Santen *et al.* 2005) such as a higher expression level of ERα, an ability to grow in low-estrogen conditions and elevated TFF1 expression. The continuous culturing of LCC1 cells in low estrogen conditions may well contribute to the increased expression of ERα in this cell line.

However, unlike most LTED-derived cells, which show little response to physiological levels of estrogen yet are sensitive to very low levels of estrogen, LCC1 cells appear truly insensitive to the addition of low levels of exogenous estrogen. Similarly, while most LTED cells show basal activation of ERK1/2 activation and ERα via Ser^{118} phosphorylation, LCC1 and

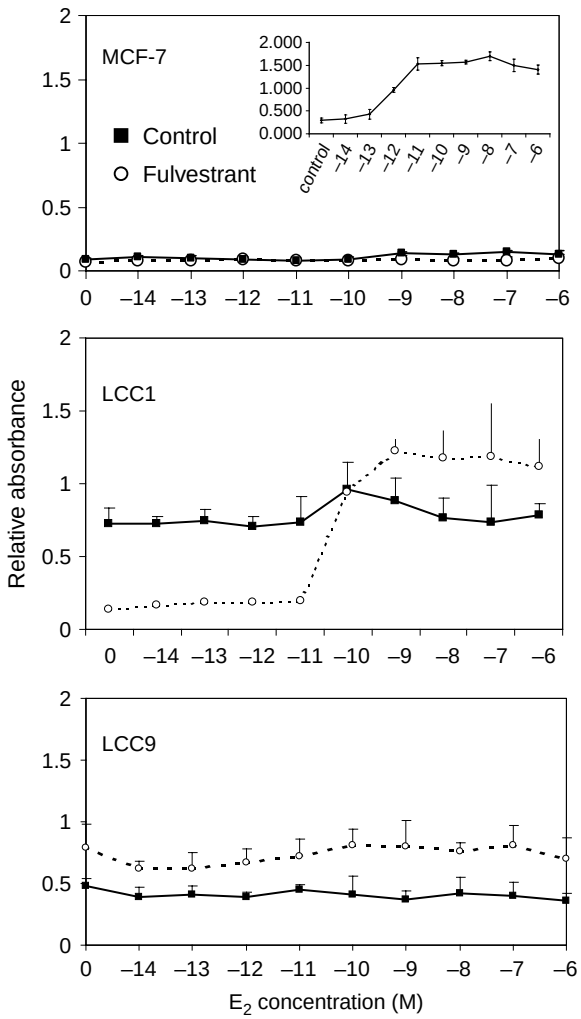


Figure 6 Effect of the growth of the cell lines in serum-free media and treated with varying concentrations of E_2 in the absence or presence of fulvestrant (1 nM). Cells were plated and after establishment placed in serum-free medium for 48 h. E_2 with or without fulvestrant was added and plates left for 72 h. Relative cell numbers were then assessed by SRB assay as described in Materials and methods. Inset in MCF-7 figure: effect of E_2 on MCF-7 cells grown in 5% double charcoal-stripped fetal serum.

LCC9 cells do not. The ER, however, is still clearly functional in LCC1 cells and linked to growth regulation as estrogen addition can produce an increase in growth which could be reversed by tamoxifen. ERα is also downregulated by the addition of estrogen and markedly phosphorylated at Ser^{118}. Additionally, the ERα downregulator fulvestrant reduces expression of TFF1 and inhibits growth. These effects are more marked when cells are exposed to fulvestrant with siRNA removal of ERα.

While constitutive activation of ERα may be achieved in some instances by phosphorylation of Ser^{118} mediated by growth factor-driven activation of ERK, an increased expression of ERα alone might account for increased DNA binding. In support of this, there was enhanced binding of ERα to the TFF1 promoter in the absence of added estrogen in both the LCC1 and LCC9 cell lines. In addition, TFF1 transcription was markedly increased in the resistant cell lines consistent with this enhanced ERα-binding driving transcription. Direct support for such a mechanism has recently been demonstrated in an MCF-7 cell line using a tetracycline-inducible ERα overexpression model (Fowler *et al.* 2004, 2006). As with the data mentioned earlier, the results suggested that elevated levels of ERα resulted in activation of receptor transcriptional function in a manner distinct from mechanisms that involve ligand binding or growth factor-induced phosphorylation of the Ser^{104}, Ser^{106} or Ser^{118} sites. The mechanism required the amino-terminal A/B domain and was not inhibited by tamoxifen. It was also uncoupled from ERK activation. The hypothesis proposed was that overexpression of unliganded ERα stabilized interactions with the basal transcriptional machinery, which at normal receptor levels may be too weak to support effective transcription (Fowler *et al.* 2004).

These results together support a model wherein growth (and TFF1 transcriptional activation) in LCC1 cells is dependent on ERα. This dependency has some ligand (i.e., estrogen) responsiveness but is largely ligand independent. The ligand-dependent component may be reversed by tamoxifen. The ligand independency appears to involve neither growth factor activation via the Ser^{118} or Ser^{167} phosphorylation routes nor hypersensitization (where low levels of estrogen produce apparent independence). Instead the ligand independence appears to be explained by the high level of ER expression leading to constitutive activation and promoting DNA binding and transcriptional activation.

We have shown that ERα is functionally active in the LCC1 model and since this has also been shown in models demonstrating LTED, a logical clinical strategy to attempt after development of resistance in a low estrogen environment (such as produced by aromatase inhibitor treatment) is to downregulate the receptor using fulvestrant (Johnston *et al.* 2005, Martin *et al.* 2005*a*,*b*). This strategy clearly is effective at inhibiting growth in LCC1 cells. However, the LCC9 variant was derived after exposure and development of resistance to fulvestrant (Brunner *et al.* 1997) and showed no growth response to either estrogen or tamoxifen. In this cell line, the negligible changes of

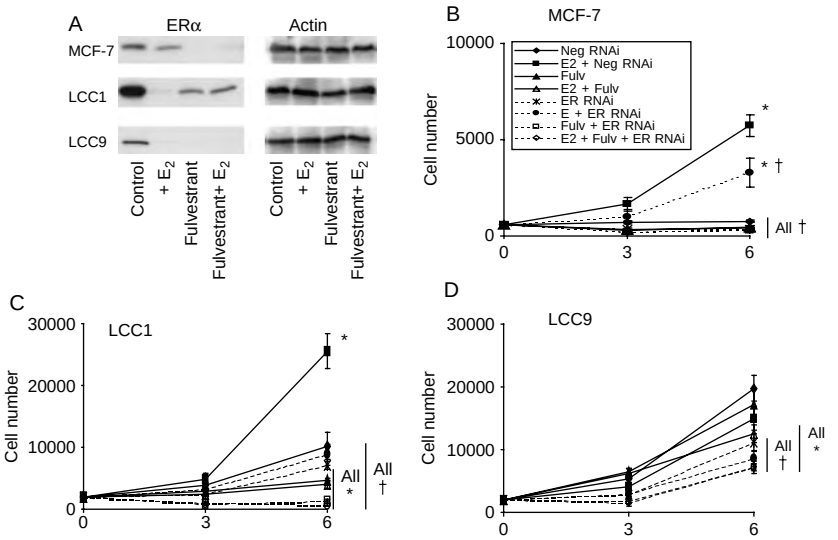


Figure 7 Effect of fulvestrant on ERα expression and combined with ERα siRNA on the growth of the cell lines. (A) Western blot analysis of ERα (66 kDa) in breast cancer cell lines in control, 1 nM E_2, 100 nM fulvestrant, or 1 nM E_2 and 100 nM fulvestrant-treated groups at 48 h. One hundred micrograms of protein were loaded per lane and detected using anti-ERα (Santa Cruz Biotech) antibody as described in Materials and Methods. Actin expression is also shown. (B–D) Effects of fulvestrant, ERα siRNA, or combinations on the growth of the cell lines. (B) MCF-7, (C) LCC1, and (D) LCC9 cells were treated with 1 nM E_2, 100 nM fulvestrant, 40 nM ERα siRNA, or combinations of these. siRNA treatment was for 4 h only, while E_2 and fulvestrant were present throughout the time course. Comparisons are made with the negative siRNA control which gave an equivalent growth effect to no treatment. Data are presented as mean + S.E. from quadruplicate samples. Statistical significance noted for treatment groups compared with matched control where $^{*}P<0.05$, negative RNAi control versus treatment group; $^{\dagger}P<0.05$, negative RNAi + E_2 control versus treatment group (ANOVA and multiple Tukey–Kramer comparison test).

ERα Ser^{118} phosphorylation obtained on estrogen or tamoxifen addition contrasted with observations in the other cell lines. Markedly reduced phosphorylation is likely to affect cofactor binding and our initial findings suggest that p160 binding (specifically AIB1) is reduced in this cell line, again consistent with endocrine insensitivity (Kuske *et al.* 2004). However, it is quite clear that fulvestrant can downregulate the receptor and even extremely high levels of fulvestrant (10 μM) were unable to influence growth (data not shown). Despite this, siRNA removal of ERα produced some growth inhibition suggesting a reduced but still measurable dependency on ERα.

In conclusion, these results suggest that multiple changes contribute to endocrine resistance. While ER still demonstrates functionality in LCC1 cells, there is a major shift to ligand independence. This independence can be explained by the high level of ER expression found in these cells and could lead to constitutive activation of the receptor. These cells still show a degree of dependency on estrogen and this can be blocked by tamoxifen. Further changes were produced by exposure and development of resistance to fulvestrant including a loss of ERα Ser^{118} activation, which could account for its loss of sensitivity to estrogen. These data support the view that in the early stages of resistance, SERDs may provide a useful therapeutic option, but other approaches will be required when resistance has developed to these agents.

Acknowledgements

The authors gratefully acknowledge support from Cancer Research UK for this study. The authors declare that there is no conflict of interest that would prejudice the impartiality of this scientific work.

References

Ali S & Coombes RC 2002 Endocrine-responsive breast cancer and strategies for combating resistance. *Nature Reviews of Cancer* **2** 101–112.

Black R, Prescott R, Bers K, Hawkins A, Stewart H & Forrest P 1983 Tumour cellularity, oestrogen receptors and prognosis in breast cancer. *Clinical Oncology* **9** 311–318.

Brunner N, Boulay V, Fojo A, Freter CE, Lippman ME & Clarke R 1993 Acquisition of hormone-independent growth in MCF-7 cells is accompanied by increased expression of estrogen-regulated genes but without detectable DNA amplifications. *Cancer Research* **53** 283–290.

Brunner N, Boysen B, Jirus S, Skaar TC, Holst-Hansen C, Lippman J, Frandsen T, Spang-Thomsen M, Fuqua S & Clarke R 1997 MCF7/LCC9: an antiestrogen-resistant variant in which acquired resistance to the steroidal antiestrogen ICI 182780 confers an early cross-resistance to the nonsteroidal antiestrogen tamoxifen. *Cancer Research* **57** 3486–3493.

Bunone G, Briand PA, Miksicek RJ & Picard D 1996 Activation of the unliganded estrogen receptor by EGF involves the MAP kinase pathway and direct phosphorylation. *EMBO Journal* **15** 2174–2183.

Clarke R, Liu MC, Bouker KB, Gu Z, Lee RY, Zhu Y, Skaar TC, Gomez B, O'Brien K & Wang Y 2001 Cellular and molecular pharmacology of antiestrogen action and resistance. *Pharmacology Reviews* **53** 25–71.

Clarke R, Liu MC, Bouker KB, Gu Z, Lee RY, Zhu Y, Skaar TC, Gomez B, O'Brien K, Wang Y *et al.* 2003 Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling. *Oncogene* **22** 7316–7339.

Dowsett M, Nicholson RI & Pietras RJ 2005 Biological characteristics of the pure antiestrogen fulvestrant: overcoming endocrine resistance. *Breast Cancer Research and Treatment* **93** S11–S18.

Fowler AM, Solodin N, Preisler-Mashek MT, Zhang P, Lee AV & Alarid ET 2004 Increases in estrogen receptor-α concentration in breast cancer cells promote serine 118/104/106-independent AF-1 transactivation and growth in the absence of estrogen. *FASEB Journal* **18** 81–93.

Fowler AM, Solodin NM, Valley CC & Alarid ET 2006 Altered target gene regulation controlled by estrogen receptor-alpha concentration. *Molecular Endocrinology* **20** 291–301.

Glass CK & Rosenfeld MG 2000 The coregulator exchange in transcriptional functions of nuclear receptors. *Genes and Development* **14** 121–141.

Howell A 2005 The future of fulvestrant (Faslodex). *Cancer Treatment Reviews* **31** S26–S33.

Jakowlew SB, Breathnach R, Jeltsch J-M, Masiakowski P & Chambon P 1984 Sequence of the pS2 mRNA induced by estrogen in the human breast cancer cell line MCF-7. *Nucleic Acid Research* **12** 2861–2877.

Johnston SRD & Dowsett M 2003 Aromatase inhibitors for breast cancer; lessons from the laboratory. *Nature Reviews of Cancer* **3** 821–831.

Johnston SR, Martin LA & Dowsett M 2005 Life following aromatase inhibitors-where now for endocrine sequencing. *Breast Cancer Research and Treatment* **93** S19–S25.

Kuske B, Moore K, MacLeod K, Naughton C, Miller WR, Smyth JF, Clarke R, Cameron DA & Langdon SP 2004 Estrogen-insensitive MCF-7 breast cancer cells show differential reduced DNA binding and estrogen receptor phosphorylation. *Breast Cancer Research and Treatment* **88** S178.

Lannigan D 2003 Estrogen receptor phosphorylation. *Steroids* **68** 1–9.

Leu Y-W, Yan PS, Fan M, Jin VX, Liu JC, Curran EM, Welshons WV, Wei SH, Davuluri RV, Plass C *et al.* 2004 Loss of estrogen receptor signaling triggers epigenetic silencing of downstream targets in breast cancer. *Cancer Research* **64** 8184–8192.

Martin MB, Franke TF, Stoica GE, Chambon P, Katzenellenbogen BS, Stoica BA, McLemore MS, Olivo SE & Stoica A 2000 A role for Akt in mediating the estrogenic functions of epidermal growth factor and insulin-like growth factor I. *Endocrinology* **141** 4503–4511.

Martin LA, Farmer I, Johnston SRD, Ali S, Marshall C & Dowsett M 2003 Enhanced estrogen receptor (ER)a, erbB2, and MAPK signal transduction pathways operate during the adaption of MCF-7 cells to long term estrogen deprivation. *Journal of Biological Chemistry* **278** 30458–30468.

Martin LA, Farmer I, Johnston SR, Ali S & Dowsett M 2005*a* Elevated ERK1/ERK2/estrogen receptor cross-talk enhances estrogen-mediated signaling during long-term estrogen deprivation. *Endocrine-Related Cancer* **12** S75–S84.

Martin LA, Pancholi S, Chan CM, Farmer I, Kimberley C, Dowsett M & Johnston SR 2005*b* The anti-oestrogen ICI 182,780, but not tamoxifen, inhibits the growth of MCF-7 breast cancer cells refractory to long-term oestrogen deprivation through down-regulation of oestrogen receptor and IGF signalling. *Endocrine-Related Cancer* **12** 1017–1036.

Masiakowski P, Breathnack R, Bloch J, Gannon F, Krust A & Chambon P 1982 Cloning of cDNA sequences of hormon-regulated genes from the MCF-7 human breast cancer cell line. *Nucleic Acid Research* **10** 7895–7903.

Metivier R, Penot G, Hubner MR, Reid G, Brand H, Kos M & Gannon F 2003 Estrogen receptor alpha directs ordered, cyclical and combinatorial recruitment of cofactors on a natural target promoter. *Cell* **115** 751–763.

Nardulli AM, Greene GL, O'Malley BW & Katzenellenbogen BS 1988 Regulation of progesterone receptor message ribonucleic acid and protein levels in MCF-7 cells by estradiol: analysis of estrogen's effect on progesterone receptor synthesis and degradation. *Endocrinology* **122** 935–944.

Nawaz Z, Lonard DM, Dennis AP, Smith CL & O'Malley BW 1999 *PNAS* **96** 1858–1862.

Nicholson RI, Hutcheson IR, Britton D, Knowlden JM, Jones HE, Harper ME, Hiscox SE, Barrow D & Gee JM 2005 Growth factor signalling networks in breast cancer and resistance to endocrine agents: new therapeutic strategies. *Journal of Steroid Biochemistry and Molecular Biology* **93** 257–262.

Normanno N, Di Maio M, De Maio E, De Luca A, de Matteis A, Giordano A, Perrone F & NCI-Naple Breast Cancer

Group 2005 Mechanisms of endocrine resistance and novel therapeutic strategies in breast cancer. *Endocrine-Related Cancer* **12** 721–747.

Parker MG 1993 Action of pure anti-estrogens in inhibiting estrogen receptor action. *Breast Cancer Research and Treatment* **26** 131–137.

Robertson JF 2002 Estrogen receptor downregulators: new antihormonal therapy for advanced breast cancer. *Clinical Therapeutics* **24** A17–A30.

Santen RJ, Song RX, Zhang Z, Kumar R, Jeng M-H, Masamura A, Lawrence J, Jr, Berstein L & Yue W 2005 Long-term estradiol deprivation in breast cancer cells up-regulates growth factor signaling and enhances estrogen sensitivity. *Endocrine-Related Cancer* **12** S61–S73.

Thorpe SM, Christensen IJ, Rasmussen BB & Rose C 1993 Short recurrence-free survival associated with high oestrogen receptor levels in the natural history of premenopausal primary breast cancer. *European Journal of Cancer* **29A** 971–977.

Yue W, Wang JP, Conaway M, Masamura S, Li Y & Santen RJ 2002 Activation of the MAPK pathway enhances sensitivity of MCF-7 breast cancer cells to the mitogenic effect of estradiol. *Endocrinology* **143** 3221–3229.

Learning the Tree of Phenotypes Using Genomic Data and VISDA

Yuanjian Feng*, Zuyi Wang†‡, Yitan Zhu*, Jianhua Xuan*, David J. Miller§,
Robert Clarke¶, Eric P. Hoffman†, and Yue Wang*
* Dept of ECE, Virginia Polytechnic Institute and State University
† Research Center for Genetic Medicine, Children's National Medical Center
‡ Dept of EECS, The Catholic University of America
§ Dept of EE, The Pennsylvania State University
¶ Lombardi Cancer Center, Georgetown University Medical Center

***Abstract*— Though supervised and unsupervised analyses of genomic data have been intensively studied in recent years, little effort has been made to discover the structural information contained in the data. In this work, we propose a stability analysis guided supervised clustering and visualization method aiming to discover the hierarchical structure in gene expression data, which we call the "tree of phenotypes". We applied the method on two multiclass gene expression microarray data sets and presented the biological plausibility of the learned trees. We also tested the multiclass classifiers built on the learned trees and demonstrated their good classification performance.**

I. INTRODUCTION

High throughput genomic data usually consists of tens of thousands of features and a relatively small number of samples of different diseases or disease subtypes. Supervised analysis of genomic data mainly refers to constructing classification schemes by learning from the dependencies between gene expression values and the given labels of phenotypes [1]. The common practice of supervised learning is to design and apply multiclass classifiers with parallel structures, such as multiclass Support Vector Machines [2], neural networks [3], and Nearest Shrunken Centroids [4]. When the number of classes increases, the complexity of a parallel-structured classifier also increases. Given limited samples in a high dimensional space, it is difficult to avoid overfitting without compromising the prediction performance. It is also hard to find a single set of genes on which the entire multiclass problem can be reliably solved.

Using tree classifiers, we can alleviate these problems [5], [6], [7]. In the tree paradigm, the original classification task is tackled by a series of simpler tasks. Each task might only require simple classification models to yield high prediction accuracy. The classification accuracy of tree-based methods is generally competitive with other classification paradigms and structures. Achieving good classification performance is not the only motivation for exploring the tree of phenotypes. The tree structure provides coarse-to-fine views of the data and its class structure, which reveals the relationship between classes (in our application context, phenotypes); moreover, at each node in the tree, we can identify a small set of features (genes) that well account for the differences between the phenotypes associated with a given node. The tree structure can be constructed based on *a priori* knowledge, or it can be learned directly from the data [5]. For example, Shedden et al. proposed a pathological tree based on tumor ontogeny [6]. Such a tree is independent of any particular data set and therefore is not affected by the sample size and the data quality. The main problem associated with this approach is that the information expressed by the tree is largely confined by the prior knowledge, which could be incomplete or inaccurate. For example in Shedden et al.'s tree, 7 tumor types, such as lung cancer and prostate cancer, are grouped together to form a single node that represents the non-Mullerian tumors. This does not provide any information about the relationships between these seven types.

In this paper, we propose a method that learns this structural information from gene expression profiles and their phenotypical labels. We first introduce the stability analysis based method for learning a tree of phenotypes from gene expression data. We then describe how to construct a tree classifier given the learned tree structure. Lastly, we apply our method on two microarray data sets, identify the biological structures obtained by our tree learning algorithm, and demonstrate the prediction performance of the tree classifiers.

II. LEARNING THE TREE OF PHENOTYPES

A tree of phenotypes is a natural way to describe the relationship between diseases or disease subtypes. We devise a method that learns the relation between phenotypes with the guidance of human interaction, namely Color-Coded Supervised Mode VIsual Statistical Data Analyzer (ccsmVISDA), an extension of the original VISDA algorithm [8]. This method gives the capability to discern unknown relationships between phenotypes that are latent within the data. To assure good generalizability for small sample sizes, the tree learning procedure includes a leave-one-out "stability" analysis that we propose. The final predicted tree is the one receiving the most votes, among all the leave-one-out trees.

A. Learning Trees by ccsmVISDA

The ccsmVISDA algorithm hierarchically displays the classes and constructs a tree. We call a tree node with two or more classes a composite node, and a tree node with only one

Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06)
0-7695-2727-2/06 $20.00 © 2006 IEEE

class a terminal node. Starting from the root node, samples are partitioned into clusters to grow the tree. A cluster is considered as a composite node if it contains more than one class; otherwise, it is a terminal node. At each composite node, the local data is first projected onto a visualization subspace that allows the user to interactively initialize the clustering. The cluster partition is iteratively updated until a stable state is reached. During the updating, samples from the same class are forced to be assigned to the same cluster, i.e. clusters learn to fully "own" either one or multiple classes.

Before constructing the tree, we first filter the genes by their signal-to-noise ratios (SNR). The purpose is to remove those non-discriminatory genes and ease the computational demand. Suppose the data set consists of K classes with p genes; each class has n_k samples, $k = 1, \ldots, K$. Denote the mean and standard deviation of data from class k and gene i by μ_{ik} and σ_{ik}, where $k = 1, \ldots, K$ and $i = 1, \ldots, p$. We define the SNR via:

$$SNR(i) = \frac{\sum_{u=1}^{K-1} \sum_{v=u+1}^{K} \pi_u \pi_v (\mu_{iu} - \mu_{iv})^2}{\sum_{k=1}^{K} \pi_k \sigma_{ik}^2} \qquad (1)$$

$$\pi_k = n_k \Big/ \sum_{j=1}^{K} n_j, \quad k = 1, \ldots, K.$$

The top m genes, with highest SNRs, are used to represent the data. Here m is proportional to K and determined empirically. We apply ccsmVISDA on the filtered data.

Suppose at a composite node there are n_0 samples with m genes coming from K_0 classes. Denote the mean vector and the covariance matrix of class c by $\boldsymbol{\mu}_c$ and $\mathbf{C}_c$. All the samples are first projected onto a two-dimensional space selected by multiclass Fisher's discriminant analysis [9], which utilizes the class information to find the most discriminatory subspace for the K_0 classes. The projection space is spanned by the two vectors that maximize Fisher's criterion [9], i.e.

$$\mathbf{W}_0 = \arg\max_{\mathbf{W}} \left\{ \mathrm{trace}(\mathbf{W}^{\mathrm{T}} \mathbf{S}_{\mathrm{w}}^{-1} \mathbf{S}_{\mathrm{b}} \mathbf{W}) \right\}, \qquad (2)$$

where $\mathbf{W}_0$ is a m by 2 matrix. Here, the within class scatter matrix $\mathbf{S}_{\mathrm{w}}$ is defined as [9]

$$\mathbf{S}_{\mathrm{W}} = \sum_{c=1}^{K_0} \pi_c \mathbf{C}_c \qquad (3)$$

and the between class scatter matrix is defined as [9]

$$\mathbf{S}_{\mathrm{b}} = \sum_{i=1}^{K_0 - 1} \sum_{j=i+1}^{K_0} \pi_i \pi_j (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}. \qquad (4)$$

Here π_c is the mixing proportion of class c, i.e.

$$\pi_c = |I_c| / n_0 \qquad (5)$$

with I_c the index set of the samples from class c, and $|I_c|$ the size of set I_c. Each sample $\mathbf{t}$ in the m-dimensional space is projected into the 2-D space

$$\mathbf{x} = \mathbf{W}_0^{\mathrm{T}} \mathbf{t}. \qquad (6)$$

Given the 2-D visualization of the samples, the user is required to determine both the number of clusters (M, $M \leq K_0$) and the initial location $\boldsymbol{\mu}_{\mathbf{x}k}$ for the center of each cluster k in the projection plot. Each cluster is modeled by a single Gaussian distribution. Denote the probability density function of a Gaussian distribution by

$$p(\mathbf{x} \,|\, \boldsymbol{\mu}, \mathbf{C}), \qquad (7)$$

$\boldsymbol{\mu}$ the mean vector and $\mathbf{C}$ the covariance matrix. In order to get a more robust partition of the samples, the user is further required to select two more partition schemes that have $M-1$ and $M+1$ clusters. All three partition schemes will undergo the same clustering process. For the partition scheme with M_0 clusters ($M_0 \in \{M-1, M, M+1\}$), after the user has selected the centers, each sample $\mathbf{x}_i$ is assigned to a cluster g_i such that

$$g_i = \arg\min_{k \in \{1, \ldots, M_0\}} \left\{ \|\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}k}\| \right\}. \qquad (8)$$

This initial partition of the data into clusters is iteratively updated by an EM-like, two step procedure. Denote the partition at the n'th iteration by

$$S^{(n)} = \left\{ S_1^{(n)}, S_2^{(n)}, \ldots, S_{M_0}^{(n)} \right\}, \qquad (9)$$

where $S_k^{(n)}$ is the index set of the samples that are assigned to cluster k.

In the first step of each iteration, each sample is assigned to a cluster. Define

$$z_{ik}^{(n)} = \begin{cases} \delta(g_i - k), & n = 0 \\ \dfrac{\pi_k^{(n)} p(\mathbf{x}_i \,|\, \boldsymbol{\mu}_{\mathbf{x}k}^{(n)}, \mathbf{C}_{\mathbf{x}k}^{(n)})}{\sum_{j=1}^{M_0} \pi_j^{(n)} p(\mathbf{x}_i \,|\, \boldsymbol{\mu}_{\mathbf{x}j}^{(n)}, \mathbf{C}_{\mathbf{x}j}^{(n)})}, & n \geq 1 \end{cases} \qquad (10)$$

where $\delta(\cdot)$ is the Kronecker delta function, i.e.

$$\delta(y) = \begin{cases} 1, & y = 0 \\ 0, & y \neq 0. \end{cases}$$

When $n \geq 1$, $z_{ik}^{(n)}$ is the *a posteriori* probability that the sample $\mathbf{x}_i$ belongs to cluster k. Each class l is assigned to a cluster k_l such that

$$k_l = \arg\max_{k \in \{1, \ldots, M_0\}} \left\{ \sum\nolimits_{i \in I_l} z_{ik}^{(n)} \right\}. \qquad (11)$$

Note that, through this operation, multiple classes may be assigned to the same cluster. The partition $S^{(n)}$ is updated accordingly:

$$S_k^{(n)} = \bigcup_{\substack{l: k_l = k \\ l \in \{1, \ldots, K_0\}}} I_l, \quad k = 1, \ldots, M_0. \qquad (12)$$

In the second step, the mean vector and covariance matrix of each cluster are updated by

$$\boldsymbol{\mu}_{\mathbf{x}k}^{(n+1)} = \frac{\sum_{i \in S_k^{(n)}} \mathbf{x}_i}{\left| S_k^{(n)} \right|}, \qquad (13)$$

$$\pi_k^{(n+1)} = \frac{\left| S_k^{(n)} \right|}{\sum_{i=1}^{M_0} \left| S_i^{(n)} \right|}, \qquad (14)$$

$$\mathbf{C}_{\mathbf{x}k}^{(n+1)} = \frac{\sum_{i \in S_k^{(n)}} (\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}k}^{(n+1)})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}k}^{(n+1)})^{\mathrm{T}}}{\left| S_k^{(n)} \right|}. \quad (15)$$

Eqs.10-15 describe the operations for updating the partition. They are repeated until the partition (Eq.9) no longer changes.

The best partition is then determined by the user, with the guidance of the Minimum Description Length (MDL) score for each of the three partitions [8]. The MDL score for the partition with M_0 clusters is

$$\mathrm{MDL}(M_0) = -\sum_{i=1}^{n_0} \log p(\mathbf{x}_i) + \frac{6M_0 - 1}{2} \cdot \log n_0, \quad (16)$$

$M_0 \in \{M-1, M, M+1\}$, where $p(\mathbf{x})$ is the probability density function of the mixture of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^{M_0} \pi_k \, p(\mathbf{x} \,|\, \boldsymbol{\mu}_{\mathbf{x}k}, \mathbf{C}_{\mathbf{x}k}). \quad (17)$$

After the optimal partition is determined, each cluster becomes a child node of the current composite node, and the partitioning is repeated until each node contains all the samples from one class.

We pursue such a human-interactive visualization approach within the mixture modeling for two complementary reasons. First, users of ccsm-VISDA can incorporate their knowledge about the relations between phenotypes during construction of the trees; second, such an approach can help users to discover new relations between phenotypes. The merits of this approach have been empirically evidenced by biological studies [10], [11].

B. The Stable Solution of the Tree

Learning the tree from all the samples may cause overfitting due to small sample size and potentially poor data quality. Thus, we embed the ccsmVISDA learning procedure within a leave-one-out stability analysis to generate "leave-one-out trees", and then take, as the final solution, the tree in this set whose hierarchical class structure has the highest frequency of occurrence. This winning tree reflects the underlying stable structural information in the data because it is learned from the data set and amongst all learned trees best survives small disturbances of the data [12]. The stability analysis based ccsmVISDA algorithm (SA-ccsmVISDA) is more robust than other solutions in the sense that, given different realizations of the data distribution, the algorithm will output similar solutions. The robustness of the tree learning algorithm is critical to making scientific discoveries, since the learned tree must be reproducible with high probability in the face of small data variations, in order to be used as a hypothesis for the underlying relations between phenotypes.

Let the set of all possible hierarchical class structures with K terminal nodes be the sample space Ω_K, which contains all distinguishable outputs of a tree learning algorithm on a given set of samples consisting of K classes. There is a distribution of the output tree structures associated with the tree learning algorithm. Let

$$T : \Omega_k \mapsto \mathbb{Z}^+, \quad (18)$$

where T is a random variable whose values are the indices of tree structures. We can use entropy to measure the stability of the tree learning algorithm, given by:

$$\mathrm{H}(L) = -\sum_{t \in \mathbb{Z}^+} p(t) \log p(t), \quad (19)$$

where $p(t)$ is the probability mass function of the derived distribution for random variable T; L is the tree learning algorithm. We can estimate the distribution $p(t)$ by the empirical distribution of the leave-one-out tree structures generated by SA-ccsmVISDA. Without loss of generality, we can always define T such that it maps the distinctive outputs of SA-ccsmVISDA to a set of consecutive positive integers starting from 1. We simply set $p(t)$ equal to 0 for those structures that do not appear in the leave-one-out trees. In the experiments we plot such empirical distributions and calculate the stability measure using Eq.19.

In the experiments, we also show the distance measures along the winning tree structure. The distance is defined as the average of the Fisher's criteria between all pairs of classes at a composite node in the 2-D projection space, and is calculated using all the samples. The distance is represented by the common length of the links between a composite node and all its child nodes.

The stability analysis based ccsm-VISDA approach for tree learning based on leave-one-out is practically feasible given the small sample sizes in most of the existing microarray data sets. The amount of human interaction required for each of the leave-one-out trials is not always the same. In the experiments we have observed that leaving one sample out each time usually will not change the structure of trees at the top levels. The projections will mostly change at the deeper levels, in the composite nodes that include the class whose samples are left out. For such instances, we only need to apply ccsm-VISDA once on those branches that are not subject to change in a specific set of leave-one-out trials. Thus, in practice, the human interaction in stability analysis based ccsm-VISDA is less intensive than that required for repetition of the ccsm-VISDA procedure over all leave-one-out training sets. If the sample sizes are increased in future microarray data sets, we can modify our approach to be semi or fully automated by exhaustively searching the optimal number of sub-clusters at each node, but in this case human users will have less or no chance to incorporate their knowledge into the structure of the trees. For existing microarray data with small sample sizes, our stability analysis based ccsm-VISDA method gives a reasonable balance between robustness and practical feasibility.

III. TREE CLASSIFIER

The tree structure learned by SA-ccsmVISDA can be used to build tree-based classifiers. Classification trees are a general

framework for solving multiclass classification problems [5]. We can use potentially any classifier as the node classifier, and use a different subspace for the particular classification task on each node. Subspace feature selection will help not only improving classification performance, but also in finding the genes that primarily account for the similarities or differences between subsets of phenotypes [6].

In the experiments, we used the feature filtering and selection method proposed by Shedden et al. [6]. First the control genes are removed from the data. Then the expression values are transformed by $\log[\max(x, 0) + 50]$. All those genes for which the standard deviation across all samples is less than 0.7 are removed. For each sub-cluster on a composite node, the set of α genes that have the highest mean expression values (and greater than the mean expression values of all the other sub-clusters combined) are selected. All these sets of genes are combined to form the subspace. Here the range of α is determined by the user and all the nodes use the same value of α. The optimal value of α is selected to achieve the lowest leave-one-out cross validation error rate of the tree classifier.

We use one-versus-rest multiclass Support Vector Machines (OVR-MSVM) [2] as the node classifiers. For a given data set, we evaluate both the hard classifier whose outputs are class labels and the soft classifier whose outputs are confidence values [13].

In the hard classification scheme, for each sub-cluster on a composite node a binary SVM is trained in the selected subspace to separate the sub-cluster from all others. A testing sample will be assigned to the sub-cluster whose associated binary SVM has the largest real-valued output, and will be passed to the child node corresponding to the sub-cluster. When the testing sample reaches a terminal node, it will be classified to the phenotype associated with the terminal node.

In the soft classification scheme, an OVR-MSVM is trained in the same way as in the hard classification scheme except that for each binary SVM the real-valued output is transformed to produce *a posteriori* probabilities. We applied the method proposed by Platt [13] to derive this transformation. The *a posteriori* probability output of an OVR-MSVM is given in the form

$$g_i^*(\mathbf{x}) = \frac{g_i(\mathbf{x})}{\sum_{k=1}^{M_0} g_k(\mathbf{x})}, \quad i = 1, \ldots, M_0, \tag{20}$$

where $g_i(\mathbf{x})$ is the transformed SVM output specified in [13]. When a testing sample is tested from the root node to each terminal node, the simulated *a posteriori* probabilities are multiplied together. The output at each terminal node is taken as the *a posteriori* probability of the sample belonging to the phenotype associated with the node. The sample is assigned to the phenotype with the highest simulated *a posteriori* probability. We will see in the experiments that the soft classification scheme improves the performance of the tree classifier.

IV. EXPERIMENTS

In the experiments, we applied the stability analysis based ccsmVISDA on the Muscular Dystrophy data set to generate the tree of 9 subtypes of muscular dystrophies and on the MIT cancer data to generate the tree of 14 cancers. For each data set, we also evaluated the performance of the tree classifiers built on the tree of phenotypes.

A. Muscular Dystrophy Data Set

The muscular dystrophy data set (provided by Children National Medical Center (CNMC), Center for Genetic Medicine) consists of 108 samples with 11252 genes from 9 diagnostic groups of muscular dystrophies. The name and the number of samples of each group are: amyotrophic lateral sclerosis (ALS, n=9); acute quadriplegic myopathy (AQM, n=5); calpain III deficiency (Calpain3, n=10); Duchenne muscular dystrophy (DMD, n=10); dysferlin deficiency (Dysferlin, n=10); fukutin related protein deficiency (FKRP, n=7); fascioscapulohumeral dystrophy (FSHD, n=14); normal human muscle (NHM, n=18); and juvenile dermatomyositis (JDM, n=25).

We applied SA-ccsmVISDA on the data and derived 108 leave-one-out trees showing 12 different structures. The empirical distribution of the trees is shown in Fig.1. The entropy calculated using Eq.19 is about 1.4208. The frequency of the winning tree is $67/108 \simeq 0.62$.

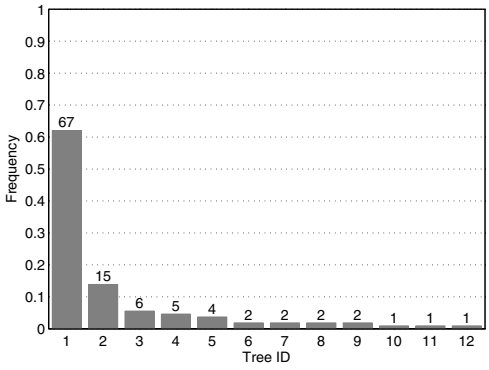


Fig. 1. The empirical distribution of the tree structures for the Muscular Dystrophy data set. The abscissa is the index of the structures, which is in descending order of frequencies. The ordinate is the frequency. The number on the top of each bar is the number of occurrences of the structure. The entropy $H(L)$ of the set of tree structures is about 1.4208.

The structure of the winning tree is shown in Fig.2 with subtype names on the terminal nodes. The winning tree is supported by many known clinical, genetic and histological features of these disorders. ALS is the only denervating disorder, due to die-back of motor neurons. A number of the muscular dystrophies are caused by abnormalities in the plasma membrane of the muscle fiber: Calpain3, DMD, Dysferlin and FKRP are all such membrane dystrophies, and all group together. FSHD is a unique disorder due to a heterozygous deletion in chromosome 4q, and by our SA-ccsmVISDA approach this maps distinctly, as does normal human muscle (NHM) and an autoimmune disease, JDM.

Based on the winning tree structure, we built hard and soft tree classifiers using the subspace selection method [6]

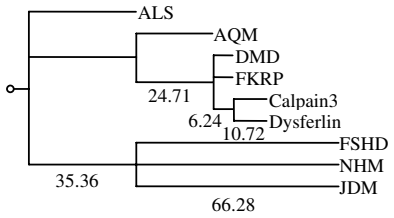


Fig. 2. The structure of the winning tree for the Muscular Dystrophy data set. The subtypes of muscular dystrophies are shown on the terminal nodes. The distance measures are shown along the links.

TABLE I

LEAVE-ONE-OUT ERROR RATES (IN PERCENTAGE) FOR HARD AND SOFT CLASSIFICATION SCHEMES ON THE MUSCULAR DYSTROPHY DATA. IN PARENTHESES ARE THE VALUES OF α. THE LOWEST ERROR IS IN BOLD FACE.

C	0.001	0.01	0.1	1.0	10.0
Hard	23.15 (800)	12.96 (250)	12.04 (50)	12.04 (50)	12.04 (50)
Soft	12.04 (300)	11.11 (400)	11.11 (300)	**10.19 (400)**	11.11 (50)

described in section 3 and OVR-MSVM as the node classifier. The gene expression values were transformed by $\log(x)$ before evaluating the classifiers. The OVR-MSVM uses linear SVMs as the binary components. The complexity of a linear SVM can be controlled by a penalty value, C. We use the same C value for each linear SVM in each of the OVR-MSVMs. We tested 5 values for C: 0.001, 0.01, 0.1, 1.0 and 10.0. In order to determine an optimal subspace size, for each C value we tested 25 different values for α: 1, 2, 5, 10, 20 and the sequence $\{50k, k = 1, \ldots, 20\}$. All these tree classifiers are evaluated by leave-one-out cross validation. In Table.I, we list the lowest error rates (in percentage) for both soft and hard classification schemes for each C value. The values in parentheses are the values of α at which the performances are achieved. In Fig.3, we plot the leave-one-out error rates of soft and hard classification as functions of α for $C = 1.0$. It can be seen that soft classification improves the performance for most cases. The lowest error rate is 10.19% when $C = 1.0$ and $\alpha = 400$.

B. MIT Cancer Data Set

The MIT cancer data was originally proposed in Ramaswamy et al. [2]. The data set has 16063 genes. It contains a training set with 144 samples and an independent testing set with 54 samples of 14 cancer types. The control genes are removed from the data before all the following experimental steps.

We applied SA-ccsmVISDA on the training set and generated 144 trees in the leave-one-out loop. The 144 trees demonstrate 20 different structures, whose empirical distribution is shown in Fig. 4. The entropy of the empirical distribution calculated using Eq.19 is about 1.3344. The frequency of the winning tree is $102/144 \simeq 0.71$. The structure of the winning tree with the cancer types shown on the terminal nodes is illustrated in Fig. 5.

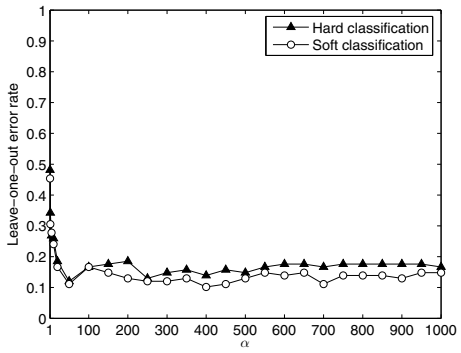


Fig. 3. Leave-one-out error rates of hard and soft classifiers on the Muscular Dystrophy data set for $C = 1.0$.

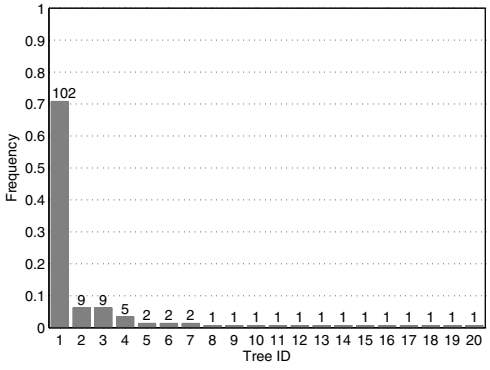


Fig. 4. The empirical distribution of the tree structures for the MIT cancer data. The entropy $H(L)$ of the set of tree structures is about 1.3344.

Using the same scheme as for the muscular dystrophy data set, we evaluated the hard and soft tree classifiers for the MIT cancer data. Before the evaluation, the gene expression values were transformed and the genes were filtered just as Shedden et al. did in [6]. The lowest leave-one-out error rates for different values of C are listed in Table.II. The curves of leave-one-out errors of soft and hard classifications as functions of α are illustrated in Fig. 6. The best classification accuracy of our tree classifiers is about 87.5% when $C = 1$ and $\alpha = 200$. Our results compare favorably with those of Ramaswamy et al. [2], who used a parallel OVR-MSVM to classify the 14 cancers and achieved a 78% accuracy on the training set of 144 samples using leave-one-out cross validation.

As a means to explore the biological implications of our solution, we compared our tree to Shedden et al.'s tree based on pathologic and ontologic knowledge [6]. Our solution has notable similarities to this tree, consistently classifying lymphoma, leukemia, CNS and epithelial cancers into groups in which lymphoma and leukemia are closely related and CNS and epithelial cancers are closely positioned. Cancers of

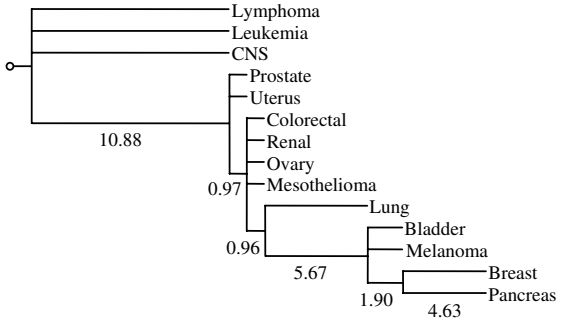


Fig. 5. The structure of the winning tree for the MIT cancer data. The cancer types are shown on the terminal nodes. The distance measures are shown along the links.

TABLE II

LEAVE-ONE-OUT ERROR RATES (IN PERCENTAGE) FOR SOFT AND HARD CLASSIFICATION SCHEMES ON THE MIT CANCER DATA. IN THE PARENTHESES ARE THE VALUES OF α. THE LOWEST ERROR IS IN BOLD FACE.

C	0.001	0.01	0.1	1.0	10.0
Hard	21.53 (450)	16.67 (200)	15.97 (200)	15.97 (200)	15.97 (200)
Soft	16.67 (250)	13.19 (250)	13.19 (200)	**12.50 (200)**	13.19 (200)

the uterus, breast, lung, colon, bladder, kidney and pancreas are consistently and appropriately classified into the same group. A major difference between the two trees is the location of the mesotheliomas (a rare cancer) and melanomas. Applying an independent data-driven approach to the same data set, Tibshirani and Hastie [7] generated a tree broadly similar to ours, in which the melanomas and mesotheliomas also clustered differently than predicted by the Shedden et al. construct. Further analysis of the similarities among the melanomas, mesotheliomas and their most closely related cancers may generate new insights into common molecular functions among these cancers.

V. CONCLUSION

In this paper, we proposed a stability analysis based data visualization algorithm that learns the tree of phenotypes from genomic data with human guidance. We applied the algorithm on two gene expression microarray data sets. The efficacy of the algorithm is demonstrated by the biological information discerned in the derived trees. We also demonstrated the prediction performance of the multiclass classifiers built on the derived trees.

This hierarchical representation of phenotypes has the power to reveal both global and local structures that are important for understanding the relationships between phenotypes. By selecting a different subspace on each composite node, we can find the genes that are important for explaining either the similarity or the difference between phenotypes in a given group. The human guided visualization approach is more robust to corrupted data and the small sample size problem than purely automated methods. By embedding the tree learning procedure

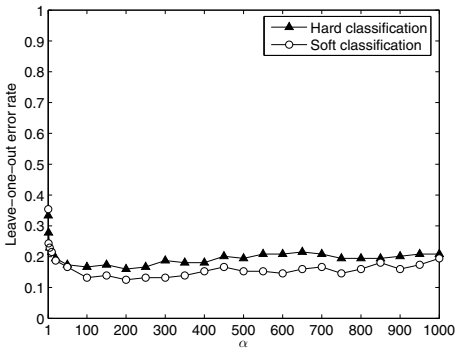


Fig. 6. Leave-one-out error rates of the hard and soft classifiers on MIT cancer data for $C = 1.0$.

within the stability analysis framework, we can generate more stable and more generalizable tree solutions given a limited number of samples.

REFERENCES

[1] G. J. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data*, Wiley, 2004

[2] S. Ramaswamy et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 26, pp.15149-15154, 2001.

[3] J. Khan et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp.673-679, 2001

[4] R. Tibshirani et al., "Diagnosis of multiple cancer types by shrunken centroids of gene expression,", *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 10, pp.6567-6572, 2002

[5] R. Simon, "Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data," *British Journal of Cancer*, vol. 89, no. 9, pp. 1599-1604, 2003.

[6] K. A. Shedden et al., "Accurate molecular classification of human cancers based on gene expression using a simple classifier with a pathological tree-based framework," *American Journal of Pathology*, vol. 163, no. 5, pp. 1985-1995, 2003.

[7] R. Tibshirani and T. Hastie, "Margin trees for high-dimensional classification," http://stat.stanford.edu/ hastie/pub, 2006

[8] Y. Wang et al., "Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 625-636, 2000.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, 1990.

[10] M. Bakay et al., "Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration," *Brain*, vol.129, pp.996-1013, 2006

[11] P. Zhao et al., "In vivo filtering of in vitro expression data reveals MyoD targets," *Comptes Rendus Biologies* , vol.326, no.10, pp.1049-1065, 2003

[12] T. Poggio et al., "General conditions for predictivity in learning theory," *Nature*, vol. 428, no. 6981, pp. 419-422, 2004.

[13] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Sch olkopf, and D. Schuurmans, Eds., pp. 61-74. MIT Press, 1999.

Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Latent Variable and nICA Modeling of Pathway Gene Module Composite

Ting Gong[1], Yitan Zhu[1], Jianhua Xuan[1, 2], Huai Li[3], Robert Clarke[4], Eric P. Hoffman[5], Yue Wang[1]
[1]Dept of ECE, Virginia Polytechnic Institute and State University, Arlington, VA, USA
[2]Dept. of EECS, the Catholic University of America, Washington, DC, USA
[3]Bioinformatics Unit, RRB, National Institutes of Health, Baltimore, MD, USA
[4]Lombardi Cancer Center, Georgetown University, Washington, DC, USA
[5]Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC, USA

Abstract – **In this paper, we report a new gene clustering approach - non-negative independent component analysis (nICA) - for microarray data analysis. Due to positive nature of molecular expressions, nICA fits better to the reality of corresponding putative biological processes. In conjunction with nICA model, VIsual Statistical Data Analyzer (VISDA) is applied to group genes into modules in the latent variable space. The experimental results show that significant enrichment of gene annotations within clusters can be obtained.**

Keywords – **non-negative ICA, latent variable model, gene clustering, module discovery, microarray data analysis**

I. Introduction

Microarray technologies provide powerful tools for genome-wide measurement of gene expressions. To discover functional modules involved in pathway signaling or gene regulation, new computational methods are required for modeling and analysis of microarray data of interest [1].

Gene clustering is widely used in the analysis of gene expression data by partitioning genes into clusters sharing similar expression patterns. The underlying assumption is that genes with similar patterns are more likely associated with common functions. Hierarchical clustering and self-organizing maps [2], have been applied to group the genes into functional modules. Recently, Independent Component Analysis (ICA) has been proposed for modeling gene clusters [3]. In contrast to traditional clustering methods, ICA-based clustering relies on a linear combination of latent biological processes and has revealed the gene clusters with significant enrichment of gene annotations or functional categories [3]. In contrast to PCA, ICA decomposes input data into components as independent as possible, showing some advantages over PCA for gene module decomposition [5].

In this paper, we report the application of non-negative ICA (nICA) for gene clustering, exploiting the non-negative nature of molecular expressions. In principle, nICA can be thought as a projection method where the expression levels are projected onto some new non-negative bases (i.e. components) with minimum statistical dependence. The nICA representation shall better reflect the biological reality. We then use VIsual Statistical Data Analyzer (VISDA) [6] to generate gene modules in the latent variable space. VISDA uses hierarchical Standard Finite Normal Mixtures (SFNM) to model clustered data where each gene belongs to each cluster with a posterior probability. The clustering procedure follows a hierarchy fasion. At each level of the hierarchy, each cluster is considered for further split, until no cluster is decomposable according to the Minimum Description Length (MDL) criterion or human justification.

This paper is organized as follows. In section II, we introduce the principle of nICA for finding gene module composites and a gradient descent algorithm of nICA. A brief description of the VISDA algorithm is also given in Section II to cluster independent components as a post-processing of nICA modeling. The application of nICA and VISDA to yeast data will be reported in Section III. Discussions and conclusions are given in Section IV.

II. Methodology

The problem of basic ICA is given according to the following linear relation:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where $\mathbf{s} = (s_1, s_2 \ldots s_N)$ is a vector of real independent sources and $\mathbf{x} = (x_1, x_2 \ldots x_N)$ is an observation vector. The assumptions of ICA are that sources are mutually independent and of non-Gaussian distribution except for at most one source. When we apply ICA to real-world problems like gene expression analysis, the situation is different from the above ideal case because of the ambiguities of ICA: the sign and permutation of sources. To resolve these ambiguities, many researchers make further assumptions to constrain the ICA model. For example, non-negativity is a natural constraint for many real-world applications, such as blind separation of natural scenes [7]. Since we assume that the underlying biological processes are independent and their expression levels should be non-negative, nICA is believed to be a more proper model to represent a linear influence of hidden cellular variables than ICA is. By projecting the data to the latent space spanned by these non-negative independent processes, fine structure of co-regulation of genes is maintained and made prominent. VISDA clustering is then applied to catch the characteristics of those subtle differences, which may lead to identify more coherent gene groups (Fig. 1).

A. nICA-based decomposition and the algorithm of nICA

As it has been known, clustering by expression pattern or "co-expressed" genes under limited experimental conditions does not provide the best possible grouping of genes by biological processes [8]. ICA-based gene clustering approach, on the other hand, is built upon a latent variable model of gene module composite. The attraction of ICA clustering lies that it can account for independent hidden effects that influence gene expression. When we introduce the non-negativity into the ICA algorithm, the resulting nICA approach can incorporate prior knowledge for better

modeling hidden sources while keeping all the advantages of ICA approach.

In our nICA model, gene express is a linear combination of biological mechanisms including pathways of signaling substances, transcription factors and their binding sites in the promoter regions of genes, as well as other different kinds of regulation [3]. We use nICA to project expression data **X** to the independent mode in order to highlight these factors. We assume that x(i, j) which is the expression level of gene i under phenotype j is expressed by the sum of non-negative independent putative biological processes $s_k(i)$, k = 1, 2, …, N, weighted by the involvement strength $a_k(j)$, k = 1, 2, …, N.

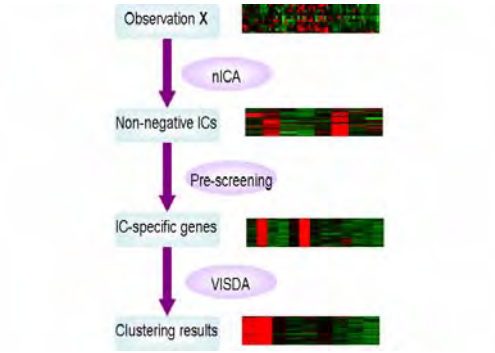


Fig. 1 Framework of nICA and VISDA for the composite module discovery

In [7], the author suggested a mathematical approach to impose non-negative constraint on sources. And if we define non-negative well-grounded sources as:

$$p(s_k < \delta) > 0 \quad \text{for } \forall \delta > 0$$
$$p(s_k < 0) = 0 \quad k = 1, 2, ..., N \tag{2}$$

then it has been proven [7] that we can find **y** = **Us**, where **U** is a square orthonormal rotation and permutation matrix, i.e. the elements of y_i of **y** are a permutation of sources if and only if all y_is are all non-negative. We note that **y** = **Us** can be re-written as **y** = **Wz** = **WVx** = **WVAs** with **V** a whitening matrix, **z** the pre-whitened observation vector and **W** an unknown orthonomal (rotation) matrix. Therefore we can consider nICA as a procedure with the following two steps: 1) remove the second order statistics by whitening; 2) search for a rotation matrix where all the data fit into the positive quadrant.

As described in [7], we can use the cost function ***J*** defined in the following to find the global minimum:

$$J(\mathbf{W}) = E\{\left\|\mathbf{z} - \mathbf{W}^T\mathbf{y}^+\right\|^2\} \tag{3}$$
$$\mathbf{y} = \mathbf{W}\mathbf{z}$$
$$y_i^+ = \max(0, y_i)$$
$$\mathbf{y}^+ = (y_1^+, y_2^+, \mathrm{L}\ \ y_N^+)$$

Based on the gradient descent rule, a learning algorithm to find the de-mixing matrix **W** is defined as follows [7]:

1) Pre-whitening the observed data **x**:

$$\mathbf{z} = \mathbf{V}\mathbf{x} \tag{4}$$

2) Using gradient descent algorithm to minimize the cost function (3):

$$\mathbf{W} = \mathbf{W} - \gamma \frac{\partial J}{\partial \mathbf{W}} \tag{5}$$

3) Projecting the unconstraint gradient descent set onto a set of orthonormal vectors.

B. Pre-screening for the clustering

After nICA, we obtain some independent components describing some distinct biological processes. In these putative biological processes, some genes showing relatively high or low expression levels are most interesting. We will use a pre-screening procedure to single out these genes.

Specifically, we can select a subsets of genes within one of the components, which includes over expressed genes (which are activated) and down expressed genes (which are repressed) according to the value of each gene in the component [3]:

a subset of genes =

$$\{genes \,|\, L_{genes} \in C\% \text{ of largest values of } y_i\}$$
$$\bigcup \{genes \,|\, L_{genes} \in C\% \text{ of smallest values of } y_i\}$$

By this pre-screening step, we actually remove some invariant genes in each component. By taking the union of the selected genes, we provide a pool of more meaningful and relevant genes to biological processes for the next step-clustering - to identify genes that belong to co-expressed modules in each component.

C. VISDA clustering

In this step, we will cluster genes into modules associated with their values in the independent components. VISDA employs the hierarchical SFNM model for hierarchical clustering. The hierarchical SFNM model uses the following probability density function to describe the relationship between successive levels in the hierarchy,

$$\mathrm{f}\left(\mathbf{r}|\boldsymbol{\pi},\boldsymbol{\theta}\right) = \sum_{k=1}^{K_0} \pi_k \sum_{j=1}^{L_k} \pi_{j|k}\, \mathrm{g}\left(\mathbf{r}|\boldsymbol{\theta}_{j|k}\right)$$
$$\sum_{k=1}^{K_0} \pi_k = 1 \quad and \quad \sum_{j=1}^{L_k} \pi_{j|k} = 1 \tag{6}$$

where **r** denotes the genes to be grouped, the upper level has K_0 clusters, the *k*th cluster in the upper level has L_K sub-clusters in the lower level, π_k is the mixing proportion of the *k*th up level cluster, $\pi_{j|k}$ is the mixing proportion of the *j*th sub-cluster in the *k*th upper level cluster, g(•) is Gaussian distribution function, $\boldsymbol{\theta}_{j|k}$ are the parameters associated with the sub-cluster. The fitting process of this model is executed by the Expectation Maximization (EM) algorithm [6], which achieves a local maximum of the likelihood function.

For each cluster at a level of the hierarchy, VISDA uses two different projection methods, Principle Component Analysis (PCA) and Principle Component Analysis – Projection Pursuit (PCA – PPM) [6], to visualize the sub-clusters within the clusters. The user chooses one of the projections that he/she thinks better revealing the data structure. On the chosen projection, user initializes models with different number of clusters by clicking on the computer screen at the centers of the clusters. These 2-D models will be

refined by EM algorithm and compete according to MDL criterion or human justification. The winning model in 2-D space will be transferred back to original data space to initialize the data model in that space. Then EM algorithm in original data space will refine the model and obtain the partition of data at that level. When no more new clusters can be found in the model validation step, the algorithm ends and a final partition is obtained.

III. Results

A. Data treatment

We applied nICA to *Saccharomyces cerevisiae* gene expression dataset [9]. The dataset contains 6152 genes with Open Reading Frames (ORFs) and 173 samples that include the different experimental conditions [9]: temperature shocks, amino acid starvation, and progression into stationary phase etc. As in [3], we also used KNNimpute to fill in missing values. And due to the triviality of clustering environmental stress response (ESR) genes defined by [9], we eliminated them in our analysis. The final dataset contains 5284 genes and 173 samples. To evaluate the experimental results, we measured the biological significance of each cluster using Gene Ontology (GO) annotation database. We mainly measured each cluster of nICA in terms of the biological process and molecular functional categories using *p*-value [3].

B. Experiment and results evaluation

To use nICA model, we took the inverse-logarithm of the data before further analysis. Hence, in our nICA model, the microarray expression corresponds to a linear additive model of interactions among biological processes. Since our goal is to find some most relevant components, dimension reduction using PCA was first applied to the data with 90% of energy maintained. Then nICA was applied to the dimension reduced dataset. As a result, we obtained the eighteen non-negative independent components. For comparison, we also did the experiment using an ICA algorithm with the same parameters.

Fig. 2 shows the first 3 independent components from nICA and ICA respectively. It is clear that, comparing to ICA, nICA is effective in separating sources as independent non-negative "biological processes" in which process-specific genes are highly biased onto two orthogonal axes respectively showed in each sub-panel in Fig. 2.

In Table 1 we only listed seven most significant clusters resulted from our nICA and VISDA approach. We measured the biological significance of each cluster using GO annotation database. The *p*-value of each cluster was calculated according to its overlap with the functional annotations in GO (see [3] for the details). Among those functional categories detected significantly by both nICA and ICA clusters, there are five out seven clusters that nICA produced significant lower *p*-values than ICA did. From these experiments, it seems to us that nICA followed by VISDA

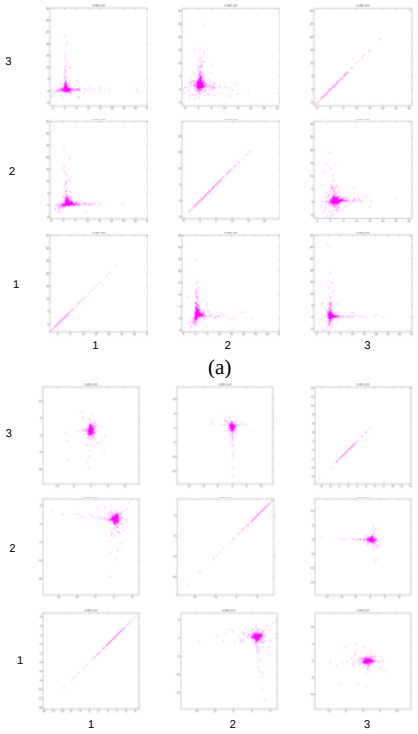


Fig. 2 Comparison of the scatter plots of the first three independent components from nICA/ICA. (a) Results from nICA; (b) Results from ICA. Each sub-panel shows the two subsequent components plotted against each other. In (a), process-specific genes are highly biased on two non-negative axes, whereas the results of ICA in (b) are not.

can extract more coherent groups of genes in terms of their functional categories.

To further evaluate our nICA-based clustering method, we used the z-score introduced in [8] to conduct a comparative study. As described in [8], the z-core is based on the mutual information between clustering results and the gene annotation. The higher scores indicate clustering results more significantly. We compared the clustering results of nICA and ICA under the same parameters and the z scores are shown in Fig. 3. As we can see from Fig. 3, nICA algorithm

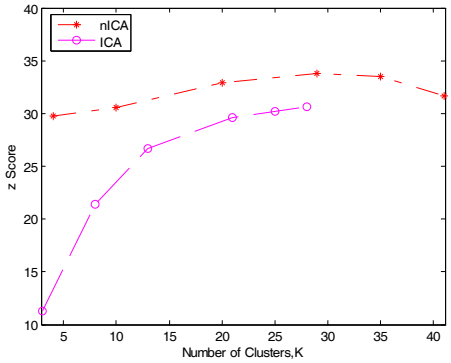


Fig. 3 Z score for the nICA (asterisk) and ICA (circle). At each level of the hierarchy of VISDA, we recorded all the intermediate clusters. At last, we got 41 clusters for nICA and 28 clusters for ICA. We inputted all these clusters to compute the z scores and drew the curves here. So based on the figures, it is reasonable to draw the conclusion that clustering methods by nICA has found a finer structure than ICA has.

TABLE I
THE SEVEN MOST SIGNIFICANT CLUSTERS OF NON-NEGATIVE ICA

Cluster index	Gene Ontology term	Cluster frequency	Genome frequency of use	P-value
6	Ty element transposition	(24 / 73, 32.8%)	(95 / 7291, 1.3%)	3.7E-27
	DNA transposition	(24 / 73, 32.8%)	(108 / 7291, 1.4%)	7.3E-26
	DNA recombination	(24 / 73, 32.8%)	(192 / 7291, 2.6%)	4.2E-20
	RNA-directed DNA polymerase activity	(14 / 73, 19.1%)	(52 / 7291, 0.7%)	2.2E-16
	DNA-directed DNA polymerase activity	(15 / 73, 20.5%)	(67 / 7291, 0.9%)	2.5E-16
16	glycolysis	(9 / 109, 8.2%)	(21 / 7291, 0.2%)	4.5E-11
	Glucose metabolism	(12 / 109, 11.0%)	(65 / 7291, 0.8%)	3.6E-10
17	proteolysis	(40 / 250, 16%)	(164 / 7291, 2.2%)	4.3E-22
	ubiquitin-dependent protein catabolism	(24 / 250, 13.6%)	(128 / 7291, 1.7%)	5.5E-20
	endopeptidase activity	(25 / 250, 10%)	(62 / 7291, 0.8%)	4.5E-19
22	amino acid and derivative/metabolism	(21 / 43, 48.8%)	(199 / 7291, 2.7%)	8.5E-22
	amino acid metabolism	(20 / 43, 46.5%)	(183 / 7291, 2.5%)	5.4E-21
	oxidoreductase	(10 / 43, 23.2%)	(247 / 7291, 3.3%)	1.4E-06
23	amino acid biosynthesis	(19 / 85, 22.3%)	(102 / 7291, 1.3%)	1.0E-17
	catalytic activity	(43 / 85, 50.5%)	(1937 / 7291, 26.5%)	2.1E-06
24	cellular response to nitrogen starvation	(4 / 47, 8.5%)	(5 / 7291, 0.0%)	3.9E-08
	cellular response to nitrogen levels	(4 / 47, 8.5%)	(5 / 7291, 0.0%)	3.9E-08
	asparagine	(4 / 47, 8.5%)	(5 / 7291, 0.0%)	3.9E-08
33	generation of precursor metabolites and energy	(32 / 68, 47.0%)	(231 / 7291, 3.1%)	8.8E-30
	oxidative phosphorylation	(19 / 68, 27.9%)	(46 / 7291, 0.6%)	4.0E-26
	hydrogen ion transporter activity	(20 / 68, 29.4%)	(55 / 7291, 0.7%)	2.1E-26

The selected clusters are listed along with the functional categories with the smallest p-value. Numbers in parentheses in the third column show the number and percentage of genes within the cluster that are presented in one of the functional category. For instance, (24/73, 32.8%) means the cluster has 73 genes, among which 24 (32.8%) genes are annotated with "Ty element transposition". And the numbers in the fourth column are presented in the similar way which corresponds to the total number within the whole genome set that are annotated with one of the special categories in GO system.

consistently performed better than ICA with an average increase of z-score of 5.

IV. CONCLUSION AND DISCUSSION

This paper presents a new gene clustering approach, namely nICA-based approach for composite module discovery. By projecting the gene expression data onto nICA space, co-regulation structure of modules are revealed and highlighted. Using a pre-screening and VISDA clustering procedure, we can identify biological process enriched clusters with coherent functional annotations. The experimental results on a yeast data set have demonstrated its advantages over conventional ICA-based approach.

Although nICA-based approach exhibits some promise for gene clustering, there is future work to be conducted. For example, we notice that in the de-mixing matrix **W**, there are some negative values that need to be properly explained, i.e., how these composite modules are involved in the corresponding biological process. Another possible direction is that we may perform gene clustering to find groups of genes under distinctive regulators or combinations of genes of these regulators.

ACKNOWLEDGMENT

This work was partially supported by the National Institutes of Health under Grants (CA109872, NS29525, EB00830, and CA096483) and the Department of Defense under Grant (BC030280).

REFERENCES

[1] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Boststein, and D. Koller, "Module networks: identifying regulatory modules and their condition-specific regulations from gene expression data," *Nature Genetics*, vol. 34, pp. 166-176, 2003.

[2] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, and E. Dmitrovsky, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 2907-2912, 1999.

[3] S.-I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, pp. R76, 2003.

[4] O. Troyanskaya, M. Cantor, O. Alter, G. Sherlock, P. Brown, D. Botstein, R. Tibshirani, T. Hastie, and R. Altman, "Missing values estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp. 520-525, 2001.

[5] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, pp. 51-60, 2002.

[6] Y. Wang, L. Luo, M. T. Freedman, and S.-Y. Kung, "Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization," *IEEE Trans on Neural Networks*, vol. 11, pp. 625-636, 2000.

[7] E. Oja and M. Plumbley, "Blind separation of positive sources by globally convergent gradient search," *Neural Computation*, vol. 16, pp. 1811-1825, 2004.

[8] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, pp. 1574-1581, 2002.

[9] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular Biology of the Cell*, vol. 11, pp. 4241-4257, 2000.